



Universidad  
Carlos III de Madrid  
[www.uc3m.es](http://www.uc3m.es)

# Grado en Ingeniería de Sistemas Audiovisuales (Plan 2011) Bilingüe

Trabajo de fin de grado

## Buscador de MOOCs orientado a VSEs

Autora:

Patricia Rodríguez Lluch

Tutor:

Antonio de Amescua Seco

**Universidad Carlos III de Madrid**

**Leganés, 2016**



# Agradecimientos

En primer lugar quiero dar las gracias a mis padres por vuestro apoyo incondicional durante toda la carrera, por haber hecho esto posible. Gracias porque me habéis convertido en la persona que soy hoy.

A mi hermana Cristina, por tu amor y tu alegría, porque haces que perseverar en conseguir tus sueños parezca algo fácil.

A Martín, por ser una fuente de inspiración con tu dedicación y excelencia en todo lo que haces, porque me ayudas a querer ser mejor. Gracias por tu ánimo y comprensión en todo momento.

Gracias a mis amigos, con los que comparto mi vida y que estáis ahí siempre, hacéis cualquier cosa más fácil. Gracias a Rafael, Andrés, Jorge y todos los demás, sabéis quienes sois.

Gracias a la gente de Agua de Vida, mi segunda familia, porque me seguís haciendo crecer cada día.

Gracias a mi tutor por su ayuda y por darme esta oportunidad.

Y gracias a todos los compañeros que han compartido estos años de ingeniería conmigo, no podría haber pedido una clase mejor ni personas más generosas.

# Abstract

Over the last years online courses have started growing in popularity, emerging new ways of learning, such as MOOC (Massive Online Open Courses). Most websites that work as MOOC aggregators platforms are focused on the student community or the teachers, being usually poorly suited for companies aiming to train their employees.

Big companies have other resources to coach their staff, in order to achieve the ETSD concept (Education, Training and Staff Development), but small companies can not keep in business having their employees occupied training and learning, as they can not afford to spare them.

This work aims to provide small companies with a web platform that can take information from the most important MOOC aggregators on the web so their employees can take online lessons in order to improve their skills.

This platform will have a feature to search for MOOC, as well as options for the business to offer recommended courses to their staff and follow their progress and will also provide their employees ways to access to courses and keep in track with them.

# Índice de contenidos

## Introducción

### Agregadores MOOC existentes y necesidades

- 2.1 Investigación de agregadores MOOC: Comparación y análisis
- 2.2 Características de un agregador MOOC orientado a VSEs
  - 2.2.1 Orientado a VSEs
  - 2.2.2 Buscador
  - 2.2.3 Benchmarking y sistema de puntuación
  - 2.2.4 Personalización
  - 2.2.5 Notificaciones y alertas
  - 2.2.6 Web Social
- 2.3 Planificación y objetivos
  - 2.3.1 Objetivos
  - 2.3.2 Planificación del trabajo
  - 2.3.3 Tabla de planificación

### Desarrollo front-end de la página web

- 3.1 Comparación de frameworks y elección
  - 3.1.1 Skeleton
  - 3.1.2 Bootstrap
  - 3.1.3 Foundation
  - 3.1.4 Elección de framework
- 3.2 Diseño de la página web y mockups
  - 3.2.1 Imagen de la empresa
  - 3.2.2 Diseño de la web
  - 3.2.3 Mockups

### Desarrollo del back-end de la página web

- 4.1 Persistencia de Datos
  - 4.1.1 Qué es la persistencia de datos y por qué es necesaria
  - 4.1.2 Gestor de sesiones
  - 4.1.3 Bases de datos
- 4.2 Ejecución del servidor

### Creación de un buscador eficiente

- 5.1 Investigación sobre creación de un buscador
  - 5.1.1 Construir una araña web
  - 5.1.2 Openwebspider
  - 5.1.3 Bobik
  - 5.1.4 OpenSearchServer
- 5.2 Desarrollo del motor de búsqueda
  - 5.2.1 Conceptos principales

- 5.2.2 Indexación
- 5.2.3 Consultas
- 5.2.4 Renderización de la página de búsqueda
- 5.2.5 Herramientas auxiliares

## **Resultados finales**

- 6.1 Página web final
- 6.2 Conclusiones
- 6.3 Trabajo futuro

# Lista de tablas

**Tabla 1.** Planificación y división de tareas

**Tabla 2.** Ventajas y desventajas de Skeleton

**Tabla 3.** Ventajas y desventajas de Bootstrap

**Tabla 4.** Ventajas y desventajas de Foundation

.

# Lista de figuras

- Figura 1:** Puntos fuertes de los diferentes agregadores MOOC evaluados.
- Figura 2:** Class central - Búsqueda de MOOC (crear alerta)
- Figura 3:** Course Talk - Sistema de puntuación
- Figura 4:** Course Talk - Editar perfil
- Figura 5:** Skeleton - Pantalla de inicio
- Figura 6:** Bootstrap - Pantalla de inicio
- Figura 7:** Foundation - Pantalla de inicio
- Figura 8:** Mockup - Pantalla de inicio de sesión
- Figura 9:** Mockup - Pantalla de búsqueda de cursos
- Figura 10:** Mockup - Pantalla de perfil de usuario
- Figura 11:** Mockup - Pantalla de perfil de empresa
- Figura 12:** Mockup - Pantalla de perfil de empresa, miembros y cursos
- Figura 13:** Navegador web solicita un documento desde el servidor web
- Figura 14:** Los datos de un envío del formulario se asignan a dos variables de sesión
- Figura 15:** Los datos previamente almacenados en una sesión se emiten
- Figura 16:** Script de cierre de sesión
- Figura 17:** Panel de control de XAMPP
- Figura 18:** Pantalla de configuración de phpMyAdmin
- Figura 19:** Openwebspider - Vista principal
- Figura 20:** Openwebspider - Documentación
- Figura 21:** Openwebspider - Modo de empleo
- Figura 22:** Bobik - Página de inicio
- Figura 23:** Bobik - Documentación
- Figura 24:** Bobik - Planes disponibles
- Figura 25:** OpenSearchServer - Inicio
- Figura 26:** OpenSearchServer - Descargas y documentación
- Figura 27:** OpenSearchServer - Documentación
- Figura 28:** OpenSearchServer - Interfaz gráfica
- Figura 29:** OpenSearchServer - Rastreador web
- Figura 30:** OpenSearchServer - Navegador de direcciones URL
- Figura 31:** OpenSearchServer - Parseadores existentes
- Figura 32:** OpenSearchServer - Atributos de los parseadores
- Figura 33:** OpenSearchServer - Mapeado de campo de los parseadores
- Figura 34:** OpenSearchServer - Editar un analizador
- Figura 35:** OpenSearchServer - Editar un clasificador
- Figura 36:** OpenSearchServer -Consultas
- Figura 37:** OpenSearchServer - Campos consultados



**Figura 38:** Plantilla de consulta

**Figura 39:** Ejemplo del uso de facetas

**Figura 40:** OpenSearchServer - Facetas

**Figura 41:** OpenSearchServer - Filtros

**Figura 42:** OpenSearchServer - Fragmentos de texto

**Figura 43:** OpenSearchServer -Mejorar subconsultas

**Figura 44:** OpenSearchServer - Autocompletado

**Figura 45:** OpenSearchServer - Renderizador

**Figura 46:** OpenSearchServer - Estilos CSS del renderizador

**Figura 47:** OpenSearchServer - Planificador

**Figura 48:** OpenSearchServer - Reproducción

**Figura 49:** OpenSearchServer - Monitorización

**Figura 50:** Página web final - Pantalla de inicio de sesión

**Figura 51:** Mockup para comparación con el resultado final

**Figura 52:** Página web final - Tablas de la base de datos

**Figura 53:** Página web final - Tabla business con los datos de empresa

**Figura 54:** Página web final - Tabla 'member' con los datos de usuario

**Figura 55:** Página web final - Pantalla de registro de usuario

**Figura 56:** Página web final - Pantalla de registro de empresa

**Figura 57:** Página web final - Formulario de registro de usuario

**Figura 58:** Página web final - Alerta de formulario de registro de usuario incompleto

**Figura 59:** Página web final - Formulario de registro de empresa

**Figura 60:** Página web final - Página inicial: Buscador de cursos y cursos en proceso

**Figura 61:** Página web final - Página inicial: Cursos en proceso y finalizados

**Figura 62:** Mockup de la página principal para comparación con el resultado final

**Figura 63:** Problemas en la pantalla de búsqueda de cursos

**Figura 64:** Página web final - Página inicial: Buscador de cursos arreglado

**Figura 65:** Página web final - Página inicial: Autocompletado en las búsquedas

**Figura 66:** Página web final - Página inicial: Resultados de búsqueda

**Figura 67:** Página web final - Perfil del usuario y cursos en proceso

**Figura 68:** Página web final - Cursos en proceso y finalizados

**Figura 69:** Mockup de la página de perfil de usuario para comparación con el resultado final

**Figura 70:** Página web final - Perfil de empresa y miembros

**Figura 71:** Página web final - Miembros de la empresa y cursos recomendados

**Figura 72:** Mockup de la página de perfil de empresa para comparación con el resultado final

# Capítulo 1

## Introducción

Nowadays the world is experiencing major changes in the way knowledge can be accessed. New technologies are changing working methods, the means by which people learn and the mechanisms to access to education [1]. Given these changes, technology and education have started complementing one another, emerging new educational models supported by digital resources in order to offer teaching and learning procedures using Information and communications technology (ICT) tools [1].

Within this framework of long distance education we can find the so-called MOOC (Massive Online Open Courses) where, unlike in classroom education which focuses on teachers, and distance education which is student-centered, MOOC learning focuses on teaching materials, often using video lectures. Since when MOOC methodology was implemented for the first time in 2007 - 2008, the process of educational innovation has immensely developed, leading to rethink the future of large-scale education and the possibility of a change in the currently known educational models. This great expansion leads to new fields where the potential possessed by MOOCs is yet to be investigated: self-learning tools even more attractive, gamification, big data analysis, identifying motivations, deficiencies and failures in the light of a massive demand. And moreover it gives us the possibility to face future risks:

Fragmentation of supply, poor reputation of accreditations, low quality, the temptation to disguise old forms of e-learning, high costs, among other risks and shortcomings [2].

In recent years it has been investigated whether MOOCs could be the solution to the problems that small and medium enterprises have in relation to the training of their employees. Companies with the aim of achieving competitiveness and increase employee performance and productivity should keep in mind the concept of "Education, Training and Staff Development" or by its acronym ETSD<sup>1</sup>. This concept is practically implemented in large companies, being a fundamental piece in its structure, while its attempts to establish in the context of small and medium enterprises (SMEs<sup>2</sup>) and especially in very small entities (VSEs<sup>3</sup>) have not been very successful. This is due to the differences between VSEs and large companies, as VSEs are not simply smaller versions of big business, but the motivations, conditions, needs, constraints and uncertainties they face are different. Due to the existence of an external uncertainty in small businesses, their objectives and goals have a shorter time perspective than in large companies, which leads them to prefer projects with short-term gains. Since ETSD is understood as a long-term project, small businesses invest less in training of its employees. [3]

---

<sup>1</sup> Education, Training and Staff Development

<sup>2</sup> Small and Medium Enterprises

<sup>3</sup> Very Small Entities

The second factor to consider is the availability of employees to begin training programs in small businesses. According to the research done by Laporte and Moreno Campos [4 - 5], VSEs have up to 25 employees, which makes them indispensable to the daily activity of the organization and activities of problem solving, usually denying the company the possibility to release their employees to conduct training sessions or other activities. Small companies are also more financially constrained than large companies, leaving little or no budget for training their employees. Besides the price of training and education programs tend to be higher for small businesses.

In 1997, Gibb [6] had already argued that certain challenges had to be defeated with the aim of bringing ETSD to small companies. The author states the need for new ways to provide training for SMEs, since the approaches that have been used to date have not given a good result in the attempt to achieve competitiveness in small businesses. Roy and Raymond [7] go deeper in this matter, saying that VSEs should make possible to their staff a continued assimilation of knowledge and skills, and thus defend the ETSD initiatives through the use of internet and web technologies

In 2013, Lujan Mora [8] evaluated the adequacy of MOOC for enterprises as a form for both, providing ETSD (i.e. training and continuing education) and supporting their staff-recruitment processes. In the same way, Hamburg *et al.* [9] state that MOOCs approach might be suitable for SMEs in order to support ETSD issues. The MOOC wave is representing right now a revolution on the e-learning scene [8]. Practically the most prestigious universities and educational institutions worldwide have enabled the massive participation of the community in a wide variety of online courses.

As Moreno Campos states [10], unfortunately today's MOOC-Aggregators do not entirely fulfill the needs of VSEs of bringing education and training to their staff. In the next chapter the author will discuss more in depth the current state of the art regarding MOOC-Aggregator Web applications, analyzing their suitability in supporting ETSD in VSEs.

# Capítulo 2

## Agregadores MOOC existentes y necesidades

En el presente capítulo se llevará a cabo una investigación de las distintas plataformas web existentes de agregadores MOOC, sus fortalezas y debilidades y los elementos necesarios para la creación de una plataforma web orientada a VSEs. El capítulo concluirá con una planificación del trabajo a realizar por la autora con el objetivo de dar forma a un agregador MOOC orientado a VSEs.

### 2.1 Investigación de agregadores MOOC: Comparación y análisis

Según el análisis previo realizado por Luján Mora [8], los agregadores MOOC más importantes en el mercado actual serían *CourseTalk*, *CourseBuffet* y *ClassCentral*. El trabajo de investigación y evaluación de agregadores MOOC llevado a cabo por Moreno Campos [10] se basa por tanto en las tres plataformas mencionadas anteriormente. A continuación se detallan los puntos más importantes concluidos por Moreno Campos en su investigación. En la figura 1 se pueden observar los principales elementos en los que se basan cada uno de los agregadores MOOC evaluados.

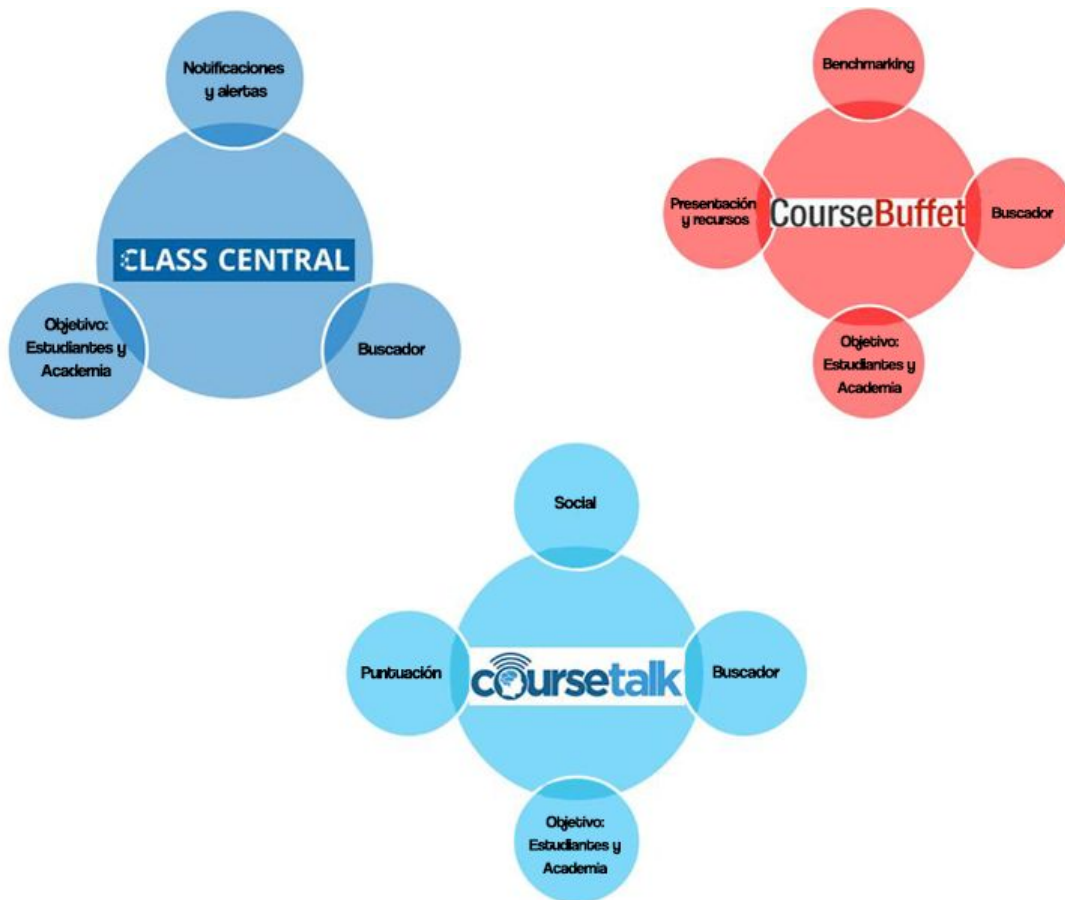


Figura 1: Puntos fuertes de los diferentes agregadores MOOC evaluados [2.1]

A pesar de algunas diferencias, Moreno Campos encuentra dos elementos coincidentes en los tres agregadores MOOC:

- Los tres agregadores MOOC están centrados en la comunidad académica, en particular para estudiantes y profesores. Esto nos muestra una falta de interés en proveer un método de enseñanza y entrenamiento del personal para SMEs, empresas pequeñas, VSEs y similares.
- Las características de búsqueda son similares en los tres tipos de agregadores, destacando la configuración de búsqueda de palabras clave que los usuarios de Class Central pueden aplicar para poner alertas para MOOCs que están anunciados o a punto de comenzar.

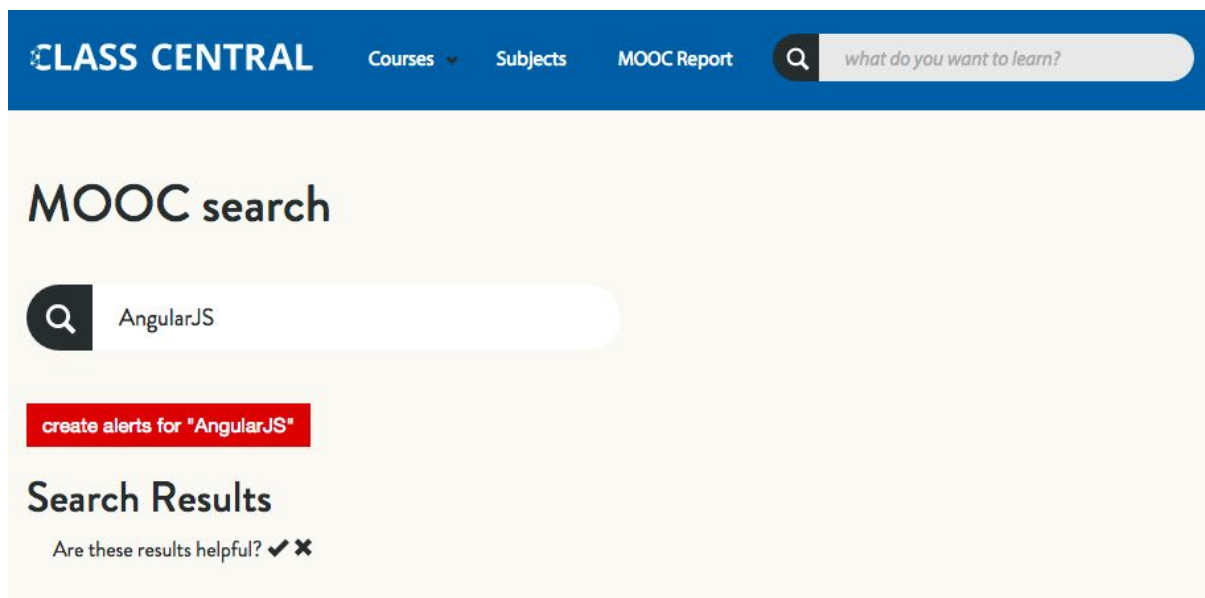


Figura 2: Class central - Búsqueda de MOOC (crear alerta) [2.1]

- En el caso de Class Central, el punto clave serían las *notificaciones y alertas* de MOOCs. Este agregador permite a los usuarios construir sus propios catálogos MOOC y recibir una notificación cuando los MOOC están a punto de comenzar o cuando vayan a ser ofrecidos de nuevo, en caso de no haber podido seguir el curso.
- En CourseTalk las fortalezas se centran en redes sociales y académicas. Para CourseTalk es de vital importancia conectar estudiantes con universidades, profesores y otros estudiantes mediante una experiencia interactiva y social para el usuario, además de conectar usuarios mediante una gran variedad de MOOCs ofrecidos por los proveedores de MOOCs más importantes. Siguiendo esta misma dirección, la web ofrece una modalidad de valoración de cursos que permite a los creadores de MOOC la posibilidad de recibir feedback real sobre sus cursos, lo cual contribuye a aumentar la calidad del MOOC ofrecido y perfeccionar la experiencia de aprendizaje en general.
- Por otro lado, CourseBuffet hace un énfasis especial en la evaluación comparativa y la agrupación. Esta web cuenta con un sistema de clasificación que ordena cada curso. Se asigna una asignatura, un nivel y un número a todos los MOOC, de manera que el usuario puede acceder al número de cualquier curso y ver qué otros MOOCs cubren la misma asignatura. Esto significa que si un usuario elige un curso online específico desde una web proveedora de MOOCs puede ser comparada su similitud con cursos MOOC ofrecidos por otras web proveedoras de MOOC. Esta web registra además las características de un MOOC específico, tales como recursos disponibles de audio, video o

libros de texto, basándose para ello en el sistema de aprendizaje del usuario. Estos recursos pueden ser utilizados por el alumno según su estilo de aprendizaje.

Estas características deben encontrarse presentes entre las características que ofrece un agregador MOOC, junto con otras propiedades, teniendo en mente el objetivo del presente trabajo de prestar apoyo ETSD en VSEs. En los siguientes puntos se hablará sobre el conjunto de propiedades que debe presentar el servicio web.

## 2.2 Características de un agregador MOOC orientado a VSEs

En el presente capítulo se hablará sobre las características que Moreno Campos identifica como necesarias para la creación de una agregador MOOC que de apoyo a VSEs.

### 2.2.1 Orientado a VSEs

Al contrario que los agregadores MOOC ya existentes, esta propuesta deberá encontrarse orientada completamente a VSEs. En el caso de los agregadores MOOC evaluados anteriormente el foco se encontraba en el sector académico, por lo que algunos de los MOOC ofrecidos no se adaptan enteramente a las necesidades de VSEs. Por ejemplo, en el caso del personal del departamento IT de una SME podría ser mejor formar parte de cursos como *“Programación C++”*, *“Fundamentos y Práctica Javascript”* o *“Construyendo una aplicación con AngularJS”* que *“Introducción a informática I”*, *“Paradigmas de la programación”* o *“Teoría de programación lineal”*. Esto dice que los MOOC ofrecidos deberían ser menos teóricos y más cercanos a los problemas que las pequeñas empresas deben afrontar cada día; es decir, cursos online orientados totalmente a las necesidades de los VSEs. Para que esta idea pudiera ser llevada a cabo en su totalidad los proveedores MOOC deberían crear más MOOCs adaptados a empresas, y a pequeñas empresas en particular.

### 2.2.2 Buscador

Un buscador eficiente debería ser el núcleo de una aplicación web para agregadores MOOC. Según Brin y Page [11] cuando presentaron Google, la intención es rastrear e indexar páginas web de la forma más eficiente posible, con la intención de producir las búsquedas de MOOC más satisfactorias para el usuario. Los autores B. An, T. Qu *et al.* [12] han notado recientemente la ausencia de servicios de búsqueda de MOOCs en los buscadores más importantes del mercado (Google, Yahoo o Bing). Informan además de una actuación muy pobre o inesperada en diferentes proveedores de MOOC como Coursera, Udacity o edX, durante la ejecución de solicitudes de búsqueda de MOOCs.

Moreno Campos [10] documenta también el haber experimentado diferentes resultados inesperados durante la actuación de solicitudes de búsqueda en las aplicaciones web de agregadores MOOC evaluadas durante el presente capítulo. Al realizar una misma búsqueda sobre un tema determinado (utilizando la misma palabra clave) en las tres páginas web de proveedores MOOC, los resultados obtenidos eran muy diferentes. An, Qu *et al.* continúan con su proposición, presentando la arquitectura de un motor de búsqueda de MOOCs vertical llamado MOOCsoso. Este buscador de MOOC cuenta con un rastreador, el cual rastrea



páginas web procedentes de distintas páginas web de proveedores de MOOCs. Estas páginas son analizadas por un analizador, en textos MOOC estructurados. Un indexador se encarga de construir un índice para recuperar los textos MOOC. En el lado del cliente, cuando los usuarios solicitan la búsqueda de un MOOC, se recuperan los resultados óptimos del texto indexado para ser representados en la interfaz gráfica de usuario. Esta arquitectura puede proporcionar una idea sobre cómo deberán ser las propiedades de búsqueda del agregador MOOC que se quiere crear en el presente trabajo.

### **2.2.3 Comparativas y sistema de puntuación**

Tal y como podemos observar en CourseBuffet, un agregador MOOC orientado a VSEs debe proporcionar mecanismos de benchmarking<sup>4</sup> que permitan a los usuarios comparar cursos online de diferentes proveedores MOOC dentro de una asignatura o tema determinado utilizando una palabra clave concreta. Esta propiedad puede facilitar a los usuarios la toma de decisiones sobre la elección de un MOOC mediante la presentación de los temas enseñados y la comparación del grado de cobertura de esos temas con otros cursos similares.

Moreno Campos [10] sugiere asimismo que todos los elementos audiovisuales ofrecidos por un MOOC específico, tales como audio, vídeo o libros de texto, sean presentados dentro de la comparación del MOOC para así facilitar el proceso de elección de un curso determinado. Igualmente, los usuarios deberían tener la oportunidad de evaluar los MOOC completados, proporcionando una crítica y sistema de puntuación tal y como podemos observar en CourseTalk (figura 3), proporcionando de esta manera un feedback real a los proveedores MOOC y los creadores de cursos y así contribuir a la mejora de MOOC, además de ayudar a otros usuarios a realizar una buena elección de MOOC basándose en puntuaciones y críticas. En la figura 3 se puede visualizar el sistema de puntuación que utiliza CourseTalk.

---

<sup>4</sup> Proceso sistemático y continuo para evaluar comparativamente productos.

## User Reviews

## Artificial Intelligence



Content ★★★★★

Instructor ☆☆☆☆☆

Provider ☆☆☆☆☆

Prior experience in the field: The AI (robotics) course on Udacity. Like: The course material was presented in an easy to absorb manner. It was engaging and you felt like learning more. The robot cartoons. They helped me stay focused - and interested! Professors' enthusiasm and knowledge. Explained stuff in easy terms. Questions - tested concepts well more often than not. Dislike: Forums weren't very interactive. Multiple choice questions can get irritating when you're not very clear with certain concepts. The projects can be very frustrating to tackle at times, but the forums usually have helpful pointers. It would be good to try Udacity's course before this one, just because it would make this a little easier. The course had a lot more work than I expected, and the work was all much more helpful in internalizing the concepts. This course is the best AI course on the web right now.

## Machine Learning



Content ★★★★★

Instructor ☆☆☆☆☆

Provider ☆☆☆☆☆

Prior experience in the field: None Like: We are introduced to a wide array of topics from basic regression to SVMs. Practical applications of the techniques was shown in large-scale projects. We got to implement what we'd learned in the lectures through some excellent (and useful) programming assignments. Dislike: The course left me feeling I had only an "overview" of machine learning, rather than being able to say I'd learned the nitty-gritty details [This could be a good thing depending on what you want]. The quizzes didn't really test much. Templates provided for every programming assignment made this course quite a bit easier than it should have been. Suggested improvements: Discard the quizzes (or make them optional). Get 1 or 2 "heavy-duty" programming assignments - no templates, you start from scratch. Overall: Good as a machine learning course, but great as an introductory course.

Figura 3: Course Talk - Sistema de puntuación [2.1]

## 2.2.4 Personalización

Un agregador MOOC necesita recoger la máxima cantidad de información sobre usuarios, tanto para los perfiles de empleados como para las empresas. Moreno Campos encuentra las propiedades observadas en CourseTalk muy adecuadas, aunque deberían ser adaptadas para los entornos de VSEs. En la figura 4 se puede observar la personalización del perfil que ofrece este agregador.

SUBJECT
PROVIDER
STARTING SOON
LEADERBOARD

PROFILE | LOG OUT

My Reviews
Edit profile

All
Your Profile
Privacy Settings
Change Email

### Your Profile

First Name

Last Name

Highest Level of Education

Masters Degree

Country

Venezuela

City

Save Profile

### Privacy Settings

Show my profile to:

Coursetalk community

Save

Figura 4: Course Talk - Editar perfil [2.1]

Un agregador MOOC orientado a VSEs debería recoger la siguiente información del usuario:

- Profesión del usuario: Cargo que ocupa el usuario, actividades realizadas y responsabilidades.
- Mayor grado académico logrado por el usuario
- Preferencias e intereses de MOOCs del usuario
- Intereses personales del usuario
- Preferencias del entorno de aprendizaje del usuario (Online / Campus)
- Tamaño de la compañía (SMEs, VSEs)
- Sector o industria de la compañía: Esta característica permitirá recibir recomendaciones de MOOC según el sector de la compañía
- Preferencias e intereses de la compañía: Áreas clave consideradas por la empresa que deben ser desarrolladas y mejoradas con el objetivo de mejorar la productividad y competitividad

Esto quiere decir que un agregador MOOC orientado a VSEs debería ser construido como un Sistema Web Adaptativo (AWS)<sup>5</sup> [13 - 14], conteniendo modelos de usuario que representen la información esencial sobre los usuarios. Es importante indicar que los aspectos de personalización pueden continuar siendo mejorados o pulidos mediante enfoques de minería de datos o webmining [14 - 15 -16]. Asimismo los VSEs deben proporcionar una interfaz donde se muestre el progreso de los miembros del personal respecto a los MOOCs en los cuales están participando.

### 2.2.5 Notificaciones y alertas

Tal y como se ha mostrado anteriormente, Class-Central dispone de un sistema de notificaciones y alertas tal y como un agregador MOOC orientado a VSEs debería proveer a sus usuarios. Este sistema de notificaciones se encarga de recordar a los usuarios cuando un MOOC deseado está a punto de comenzar, así como de crear alertas conforme a búsquedas de ciertas palabras clave. Un aspecto que no se encuentra presente en Class-Central pero puede ser interesante implementar es una interfaz VSE (como se comenta en la sección 2.4) para recomendar a miembros del personal participar en cursos mediante la creación de alertas de los MOOC disponibles y deseados y avisar sobre esas alertas o recomendaciones de MOOC. De esta manera, se anima a los empleados a unirse a ciertos cursos considerados “MOOC clave” por la empresa.

### 2.2.6 Web Social

Conforme a las tendencias actuales, es prácticamente imposible imaginar una aplicación Web sin componentes sociales. Appelquist *et al.* [17] definen la Web Social como un grupo de relaciones que unen diferentes personas alrededor de Internet, lo cual significa que la Web Social se extiende alrededor de internet. CourseTalk ofrece aspectos sociales muy interesantes que pueden ser considerados, con algunas modificaciones, para un agregador MOOC orientado a VSEs. El agregador debería conectar usuarios con otros usuarios mediante una red social interactiva:

- Los empleados podrían compartir experiencias, ideas y conocimientos con otros empleados realizando los mismos cursos.
- Las VSEs podrían comunicarse con otras pequeñas empresas para intercambiar experiencias ETSD

---

<sup>5</sup> Adaptive Web System

- Las compañías obtienen una oportunidad de contactar con posibles empleados para cubrir algunos puestos y, de esa manera, reforzar su personal y proceso de contratación.

El agregador debería además ser capaz de unir los perfiles de usuario de sus redes sociales profesionales (como LinkedIn o Xing<sup>6</sup>) con sus respectivas cuentas del agregador MOOC.

---

<sup>6</sup> Xing es la red social profesional más grande de los países Europeos de habla Alemana, con 14 millones de miembros y ofrece contactos profesionales entre empleados y empresarios.

## 2.3 Planificación y objetivos

Siguiendo el análisis de los agregadores MOOCs ya vistos en los puntos anteriores, se presenta una planificación para el trabajo a realizar, eligiendo unos puntos clave en los que centrarse de los necesarios que posea un agregador MOOC orientado a VSEs, expuestos en el punto 2.2 de este mismo trabajo.

### 2.3.1 Objetivos

La autora del trabajo ha optado por centrarse en tres de los puntos recogidos anteriormente:

- Orientado a VSEs: Este es el punto vital para el presente trabajo, ya que es lo que diferencia el buscador de MOOCs que se quiere crear del resto del mercado.
- Buscador: Al finalizar el proyecto se deberá disponer de un buscador funcional que pueda encontrar cursos MOOC de distintos proveedores al proveer una palabra clave.
- Personalización: Se crearán perfiles personalizados según el tipo de usuario, con datos guardados en bases de datos, como se desarrollará más adelante.

### 2.3.2 Planificación del trabajo

El proyecto será realizado con la siguiente estructura:

- En primer lugar, será necesaria la creación de una página web. El primer trabajo a realizar por la autora será una búsqueda de los principales frameworks<sup>7</sup> adaptativos del mercado y la elección del que más se adapte a las necesidades de un agregador MOOC orientado a VSEs.
- Una vez elegido, será necesario diseñar la web. Primero se realizará un mockup o modelo de la parte front-end de la página web para tener un diseño base sobre el que trabajar, a partir de ello, se programará la página web utilizando tecnologías HTML5, CSS, Javascript y PHP.
- Posteriormente, se creará una base de datos para recoger los elementos que componen el perfil de usuario.

---

<sup>7</sup> Estructura conceptual y tecnológica de soporte definido que puede servir de base para la organización y desarrollo de software

- Se utilizarán dos estilos de perfil diferentes según se trate de un empleado o una empresa, tal y como se expone en el punto 2.2.4: Personalización, y se iniciará sesión de manera independiente para cada tipo de usuario.
- Finalmente, una vez realizada la parte front-end de la web se investigará el asunto de realizar un buscador para la página que reciba información de los principales proveedores de MOOCs y que devuelva al usuario información de los cursos relevantes al buscar una palabra clave. Este punto será la tarea clave del trabajo y ocupará la mayor parte del tiempo del proyecto.

### 2.3.3 Tabla de planificación

Este apartado consiste en una tabla con el esquema de la planificación del tiempo que se utilizará para cada actividad.

Tabla 1. Planificación y división de tareas

<i>Actividad</i>	<i>Descripción</i>	<i>Fecha inicio</i>	<i>Fecha Final</i>
Investigación MOOC	Investigación de los MOOC existentes y lectura de artículos	29/02/2016	7/03/2016
Investigación de frameworks y	Estudio de frameworks disponibles, realizando un documento donde se expliquen sus características	7/03/2016	14/03/2016
Decisión elementos web	Decisión sobre las características y elementos que tendrá la página creada	14/03/2016	21/03/2016
Mockup	Diseño de la web y realización de mockups	21/03/2016	11/04/2016
Investigación BBDD	Investigar sobre bases de datos y back-end	11/04/2016	28/04/2016
Entorno de desarrollo	Encontrar un servicio de servidor sobre el que se pueda sostener la página web	28/04/2016	02/05/2016
Web	Realización de la parte front-end del buscador	02/05/2016	16/05/2016

BBDD	Crear la base de datos para la web, que recoja la información del usuario para que pueda iniciar sesión y se muestre en el perfil	16/05/2016	30/05/2016
Orientación a empresas	Crear el perfil del usuario y perfil de empresa, añadir la parte de empresa a la base de datos	13/06/2016	27/06/2016
Investigación para crear el buscador	Investigación sobre el concepto de web crawling para buscar cursos de forma eficiente orientado a VSEs	27/06/2016	11/07/2016
Web crawling	Creación de un web crawler	11/07/2016	25/07/2016

---

Esfuerzo dedicado al trabajo (Número de horas aproximadas): 500 horas



# Capítulo 3

## Desarrollo front-end de la página web

### 3.1 Comparación de frameworks y elección

Para realizar la interfaz de la página web o front-end se ha optado por el empleo de un framework como base para el desarrollo.

Un framework CSS o Front-End consiste generalmente en un paquete compuesto por una estructura de ficheros y carpetas de código estandarizado (Documentos HTML, CSS, Javascript...) [18].

Los componentes habituales de un framework completo son:

- Código fuente CSS para crear una cuadrícula: Esto permite al desarrollador posicionar los diferentes elementos que componen el diseño de la página de manera sencilla y versátil.
- Estilos de tipografías definidos para los elementos HTML.
- Soluciones para los casos de incompatibilidad entre navegadores, de manera que la página se muestre de forma correcta en todos los navegadores.
- Creación de clases CSS estándar que pueden ser utilizadas para estilizar componentes avanzados de la interfaz de usuario.

En la búsqueda de los frameworks más relevantes del mercado la autora ha escogido los siguientes frameworks adaptativos: *Bootstrap*, *Foundation* y *Skeleton*.

#### 3.1.1 Skeleton

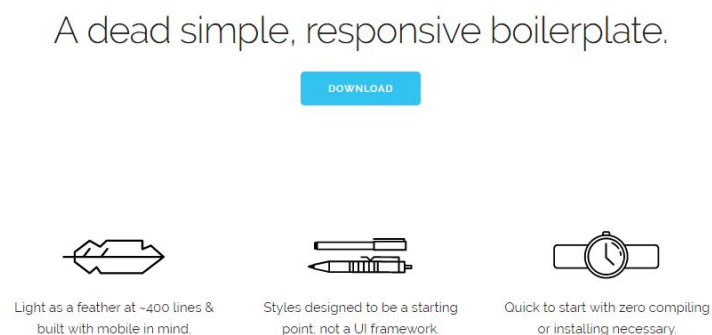


Figura 5: Skeleton - Pantalla de inicio

Skeleton [19] se trata de un CSS simple y adaptativo para páginas web HTML5 y aplicaciones. [20 - 21]

Es ligero y más eficiente para sitios pequeños, ligero y preciso. Se enfoca primeramente en navegación y tipografía y es muy fácil comenzar a utilizarlo. [22]

Tabla 2. Ventajas y desventajas de Skeleton

<i><b>Ventajas</b></i>	<i><b>Desventajas</b></i>
Fácil de aprender	Incompleta en solitario, mejor para combinar con otros frameworks
Cuadrícula de diseño adaptativa	Para páginas web demasiado simples
Clase CSS para elementos de imagen adaptativos que escalan con la cuadrícula de diseño	No se pueden anidar columnas
Plantilla PSD para realizar mockups	Pocos estilos
HTML shiv, para que HTML5 sea correctamente interpretado en navegadores antiguos	
Ligero	

### 3.1.2 Bootstrap

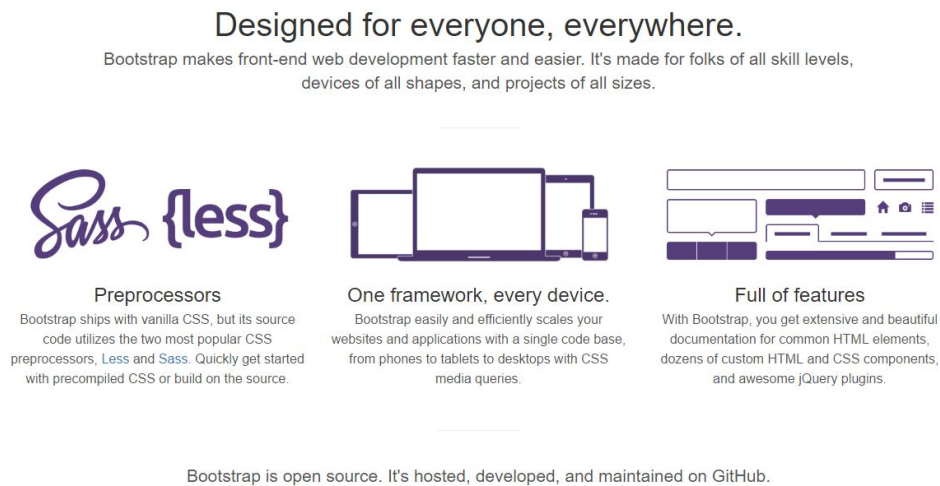


Figura 6: Bootstrap - Pantalla de inicio

Bootstrap [23] es un framework de desarrollo front-end moderno y muy popular. Contiene una gran cantidad de funciones y es un framework muy completo para desarrollar páginas web adaptativas y aplicaciones [20].

Tabla 3. Ventajas y desventajas de Bootstrap

<i><b>Ventajas</b></i>	<i><b>Desventajas</b></i>
Rápido	No sigue las mejores prácticas: hace que la escalabilidad, la reutilización y el mantenimiento sean más difíciles
Bien documentado	Pesado
Fácil de utilizar	Las páginas creadas con bootstrap son muy parecidas, es poco original
Para proyectos de cualquier tamaño	No funciona con javascript deshabilitado
Diseño de cuadrícula fluida	Difícil agregar o cambiar contenido una vez configurado
Diseño adaptable	
Totalmente personalizable	

### 3.1.3 Foundation

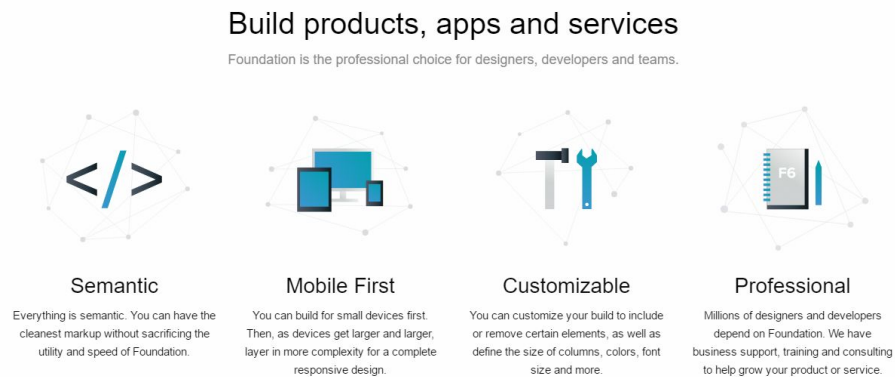


Figura 7: Foundation - Pantalla de inicio

Foundation [24] es otro framework front-end adaptable que goza de una gran popularidad. Con foundation es posible diseñar páginas web tanto para comenzar el desarrollo pensando en pantallas de teléfonos móviles como para pasar de grandes pantallas a tamaños móviles [20]. Foundation tiene su propio centro de entrenamiento y videos para una experiencia de aprendizaje impecable [22].

Además, cuenta con una web que contiene distintos elementos mostrando su código para poder añadir de manera sencilla diferentes funciones a la web.

Tabla 4. Ventajas y desventajas de Foundation

<i><b>Ventajas</b></i>	<i><b>Desventajas</b></i>
Sistema de cuadrículas más robusto	Como en Bootstrap, existe el peligro de crear un sitio web de aspecto genérico, pero Foundation es más fácil de personalizar después de configurarlo, así que esto no es tanto problema
Vídeo tutoriales	El desarrollo en Bootstrap es más rápido
Está hecho para desarrolladores y diseñadores: si se quieren agregar más cosas o hacer algunos cambios es más fácil	
Es muy liviano: no es sólo es adaptable, si no que está optimizado para móviles	
Mejores plugins	

### **3.1.4 Elección de framework**

Después de realizar la comparación de los tres frameworks y según los criterios evaluados la autora ha tomado la decisión de optar por Foundation, ya que tiene un sistema de cuadrículas más robusto y ofrece la posibilidad de personalizar más la página web, dando un aspecto más profesional y diferenciado al resultado final.

Además, Foundation tiene una documentación excepcional, con su propio centro de entrenamiento y un sistema de enseñanza mediante video tutoriales.

## **3.2 Diseño de la página web y mockups**

Previamente a comenzar la programación de la página web se ha de determinar la imagen que se quiere dar a la página web conforme al servicio que va a realizar.

### **3.2.1 Imagen de la empresa**

La imagen que se quiere ofrecer al cliente es una empresa seria, con un componente social orientado al trabajo (estilo LinkedIn), teniendo como objetivo empresas pequeñas y medianas y enfocada en mejorar las habilidades de los miembros de la empresa.

Los colores elegidos para el diseño de la página web son el azul y el naranja, ya que el azul aporta seriedad y profesionalidad y el naranja es un color dinámico y en sintonía con la idea de aprendizaje, realización de cursos y la parte social de la web.

### **3.2.2 Diseño de la web**

Como se ha mencionado en el apartado de Planificación (2.3), se mostrará una página distinta según si el usuario inicia sesión como empleado o como empresa.

La versión para empleados se centrará en la búsqueda de cursos online y en el perfil de usuario, donde se podrán ver los MOOC realizados y los MOOC en curso. El perfil de usuario contendrá la información especificada en el apartado de Personalización (2.2.4):

- Profesión del usuario: Cargo que ocupa el usuario, actividades realizadas y responsabilidades.
- Mayor grado académico logrado por el usuario.
- Preferencias e intereses de MOOCs del usuario.
- Intereses personales del usuario.
- Preferencias del entorno de aprendizaje del usuario (Online / Campus).

La versión para la empresa consistirá en un perfil que incluirá los siguientes datos sobre la compañía:

- Tamaño de la compañía (SMEs, VSEs).
- Sector o industria de la compañía: Esta característica permitirá recibir recomendaciones de MOOC según el sector de la compañía.
- Preferencias e intereses de la compañía: Áreas consideradas clave por la empresa que deben ser desarrolladas y mejoradas con el objetivo de mejorar la productividad y competitividad.

El perfil de empresa contendrá un apartado con los perfiles de los empleados, donde se podrá observar el progreso en los cursos de los miembros y además permitirá la posibilidad de agregar cursos recomendados para sus empleados.

### **3.2.3 Mockups**

Para realizar los modelos para el diseño que van a componer la página web la autora ha utilizado el programa para edición de imágenes Adobe Photoshop [25].

En primer lugar, se ha realizado el diseño de la pantalla de inicio de sesión que se encuentra al ingresar en la dirección web por primera vez (figura 8). Como se ha comentado anteriormente, se dispone de dos maneras diferentes para realizar el inicio de sesión, según se trate de un empleado o un responsable de la empresa, ya que el contenido de la página web será diferente en cada caso. Se ha optado por utilizar una barra de color gris oscuro como encabezado, que contendrá el logo o nombre de la aplicación web en el lado izquierdo, y las pestañas para navegación por la página a la derecha después de haber iniciado sesión.

El color naranja se ha utilizado para la parte de los trabajadores, ya que aporta más dinamismo y vitalidad. El azul aporta seriedad y profesionalidad para la parte de empresas. En la parte inferior de cada formulario de inicio de sesión se encuentran los enlaces para recuperar la contraseña en caso de pérdida y el registro para crear una nueva cuenta, este enlace llevará a un formulario distinto según se quiera crear cuenta como empresa o como empleado.

Idealmente, la sesión se podrá guardar entre conexiones y no será necesario iniciar sesión cada vez que el usuario quiera conectarse a la página web.



Figura 8:Mockup - Pantalla de inicio de sesión

A continuación podemos observar en la figura 9 la página que se encuentra el empleado una vez ha iniciado sesión. Esta será la página de inicio o home que se cargará cada vez que el usuario se conecte.

Esta página está compuesta por:

- Un buscador donde se puede introducir una palabra clave o asignatura para buscar nuevos cursos a los que apuntarse. Al pulsar el botón “Buscar” se redirigirá al usuario a una nueva ventana con los resultados de la búsqueda.
- Un slider donde el usuario podrá visualizar los cursos que está realizando actualmente y el progreso completado del curso. El trabajador podrá acceder al curso en cuestión haciendo clic en el mismo.
- Un slider donde se podrán observar los cursos ya finalizados por el usuario. Se podrá acceder a ellos haciendo clic en el curso finalizado.





Figura 9: Mockup - Pantalla de búsqueda de cursos

Al pulsar la pestaña “Perfil” en la barra de navegación superior el empleado podrá acceder a su perfil y realizar cambios en el mismo (figura 10). Esta vista se compone por los siguientes elementos:

- Un apartado donde se muestra la información personal del usuario, con todos los datos especificados anteriormente. El usuario podrá editar esta información desde esa misma ventana.
- Notificaciones y Alertas: Pese a que el desarrollo de estas dos características queda fuera de la extensión de este trabajo, la autora ha optado por introducirlas en el diseño del modelo como referencia para futuras mejoras de la página web.
- Un slider donde el usuario podrá visualizar los cursos que está realizando actualmente y el progreso completado del curso. El empleado podrá acceder al curso en cuestión haciendo clic en el mismo.

- Un slider donde se podrán observar los cursos ya finalizados por el usuario. Se podrá acceder a ellos haciendo clic en el curso finalizado.

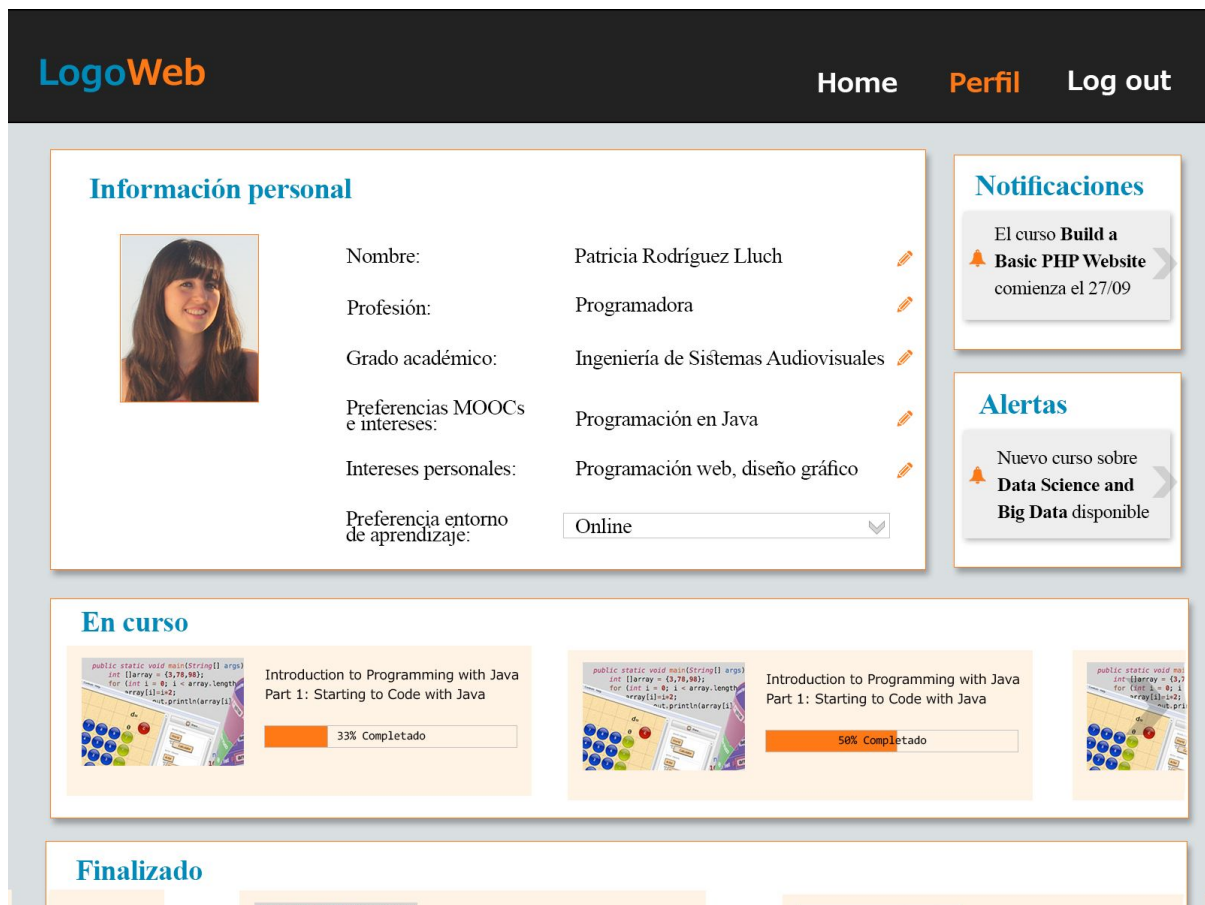


Figura 10: Mockup - Pantalla de perfil de usuario

Finalmente, en la figura 11 se puede observar la pantalla que se encuentran los responsables de la empresa al iniciar sesión. En esta vista se encuentra:

- Un apartado con la información de la empresa que se quiere mostrar, la cual ha sido especificada en el apartado anterior. El responsable de administrar el perfil de la empresa podrá modificar los datos de la empresa desde esta misma pestaña.
- Una opción para agregar nuevos miembros a la empresa
- Una opción para agregar cursos que la empresa considera importantes para que sus empleados realicen.

Al acceder a este apartado se abrirá una ventana emergente, donde se encontrará el buscador para introducir una palabra clave. Al realizar esta acción se accede a la misma vista que

utilizando el buscador de la figura 9, desde esa vista el trabajador podrá seleccionar el curso deseado.

- Un slider donde se visualizan los empleados que se encuentran registrados en la página web. Se mostrará la información de uno de los curso que se encuentren realizando o al que estén inscritos y el progreso en el mismo. Se podrá acceder a los perfiles de los empleados haciendo clic sobre la imagen o descripción.

El perfil que se visualiza al realizar esta acción será similar al mostrado en la figura 10, pero en esta ocasión no se podrá editar la información de perfil y no se hallarán los apartados de alertas y notificaciones.

- Un slider con los cursos recomendados agregados por la empresa, con opción de acceder al curso para visualizarlo haciendo un clic sobre el mismo.



Figura 11: Mockup - Pantalla de perfil de empresa



Figura 12: Mockup - Pantalla de perfil de empresa, miembros y cursos

# Capítulo 4

## Desarrollo del back-end de la página web

En este capítulo se hablará de la labor que compone el acceso a bases de datos y la implementación de características vitales para el funcionamiento de la página web.

### 4.1 Persistencia de Datos

#### 4.1.1 Qué es la persistencia de datos y por qué es necesaria

Para que el buscador que se quiere desarrollar no consista únicamente en una página web estática, sino que pueda recoger información a través de la web, se necesita crear una capa de acceso a datos [26].

Utilizando una base de datos conjuntamente con un servidor web se puede almacenar información estructurada de manera similar (tal como datos de clientes, respuestas a encuestas o información de contacto) en un único lugar, y de esta manera producir de forma automática numerosas páginas web sobre la marcha, como un directorio de empleados o una aplicación de solicitud de servicios. Se puede reutilizar fácilmente determinada información en determinadas situaciones, (por ejemplo, mantener una lista con todos los empleados junto a páginas de inicio separadas para cada empleado) necesitando poco mantenimiento para ello. Una vez que el sitio web necesite un rediseño solo habrá que modificar un archivo de plantilla en lugar de un gran número de páginas.

Las aplicaciones web son por naturaleza aplicaciones distribuidas. Esto significa que hay una parte de la aplicación que se ejecuta en el servidor web, mientras que otra parte se ejecuta en el equipo cliente dentro de una ventana del navegador web (ver figura 13).

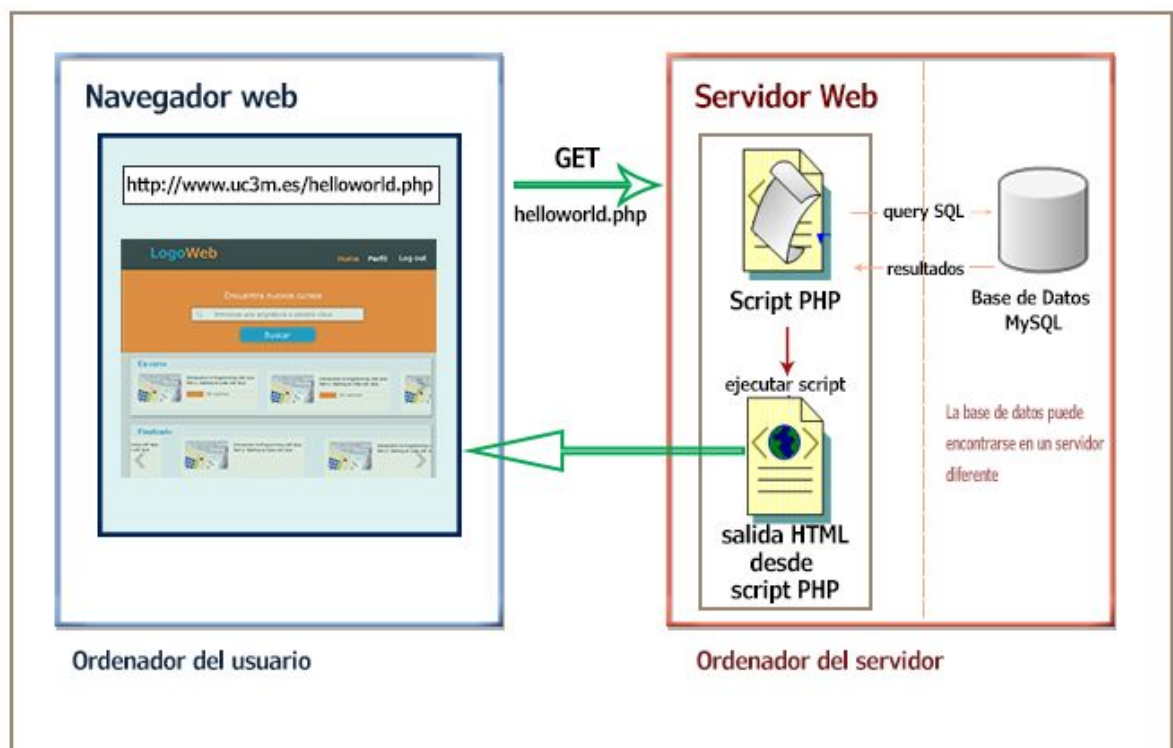


Figura 13: Navegador web solicita un documento desde el servidor web

Cuando el usuario escribe una dirección web (URL = Uniform Resource Locator) y pulsa la tecla enter (o al hacer clic en un enlace o elige un marcador), el navegador web envía esta dirección en una manera denominada como solicitud HTTP (Del inglés, HyperText Transfer Protocol) al servidor web. Si el servidor web recibe una petición de un archivo HTML estático (normalmente indicado por un nombre de archivo con extensión .html o .htm) simplemente localiza el archivo correspondiente en su disco duro y lo envía de vuelta al equipo que lo había solicitado. Si el servidor web recibe una solicitud de archivos dinámicos (a menudo indicado por extensiones .php, .asp, .cfm, .jsp, etc.), realiza algunas tareas extra antes de enviar una respuesta al navegador web.

El navegador web se encarga de mostrar el diseño descrito por el archivo HTML que ha recibido (este proceso se denomina "renderización de páginas"). Una vez que la página web se representa en la pantalla el usuario puede interactuar con él (hacer clic en los enlaces, rellenar formularios, etc). Dado que HTML proporciona sólo un nivel de interactividad básico (es decir, enlaces, elementos de formularios, botones y mapas de imagen), se creó JavaScript. JavaScript es un lenguaje de programación que da al diseñador de páginas web una serie de opciones para la creación de interfaces de usuario sofisticadas (por ejemplo, validación de entrada, ventanas pop-up, menús en cascada). El código JavaScript a menudo se contiene dentro de un archivo HTML utilizando la etiqueta `<script>`. JavaScript se ejecuta siempre por el navegador web. El servidor web es totalmente ajeno a JavaScript; sólo muestra

todo el código HTML, JavaScript o imágenes que se han solicitado sin analizar o ejecutar cualquier código.

Si se quieren construir las páginas web a partir de los datos contenidos en una base de datos, el servidor web debe trabajar un poco más. Si el servidor web recibe una petición de un archivo dinámico (indicado por un archivo que termina .php y algunos otros) el servidor web tendrá que localizar el archivo en su disco duro e interpretar (también denominado parsear) y ejecutar su contenido. PHP (acrónimo recursivo de "PHP: Hypertext Preprocessor") es un lenguaje de programación que se utiliza para la creación de páginas web dinámicas y siempre se ejecuta en el servidor web (en contraposición al navegador web en el ordenador del cliente). Generalmente los comandos PHP producen determinado código HTML en función de la entrada del usuario o los datos leídos de una base de datos, etc. Este código HTML resultante se envía de vuelta al navegador web para su renderizado.

HTTP (Protocolo de Transferencia de Hipertexto) "tiene mala memoria" (es un llamado protocolo sin estado). Cuando un script PHP recibe datos de un formulario HTML, esta información sólo está disponible durante la ejecución del script. Esto significa que por defecto sólo se pueden procesar los datos inmediatamente después de que se ha enviado, pero no más tarde. Obviamente, esto causa problemas, por ejemplo cuando se desea que la aplicación recuerde las credenciales de inicio de sesión de un usuario durante de una sesión completa. Hay varias maneras de abordar este problema:

- Utilizar variables de formulario HTML ocultos: Esta solución conlleva muchos riesgos para la seguridad de una aplicación web, ya que puede ser falsificada fácilmente.
- Utilizar la gestión de sesiones (a través de cookies y URL).
- Utilizar una base de datos.

Hoy en día, una aplicación web normalmente utiliza la gestión de sesiones para la persistencia de datos a corto plazo y bases de datos para la persistencia a largo plazo.

### **4.1.2 Gestor de sesiones**

El tiempo durante el cual una persona usa un sitio web se denomina sesión. Con bastante frecuencia, se quiere que el servidor web recuerde cosas que sucedieron en el pasado (como el inicio de sesión de un usuario, un envío de formularios, etc.) de manera que se puedan reutilizar datos sin tener que pedir al usuario que vuelva a introducir la información todo el tiempo. Por desgracia, hay que decirle al servidor web de forma explícita qué datos se debe recordar. Esto se realiza de la manera que se muestra en la figura 14 y la figura 15. En la Figura 14 un PHP script recibe los datos (nombre y apellido) de un formulario HTML y lo almacena en dos de las llamadas variables de sesión, es decir, `$_SESSION["nombre"]` y `$_SESSION["apellido"]`. En la figura 15 otro script PHP tiene acceso a la información

mediante la lectura de las variables de sesión. Esto puede ocurrir 10 minutos más tarde y en cualquier script PHP. El único requisito para que funcione la gestión de sesiones es que se llame a la función `session_start()`; al principio de cada script PHP que utiliza la información de sesión. La información almacenada en la variable `$_SESSION` está disponible para la duración de toda la sesión. Sin embargo, la sesión puede caducar después de un cierto período si el navegador no ha comunicado los datos al servidor web, es decir, el usuario se encontraba inactivo.

```
<?php
    session_start();
?>

...

<?php
    $_SESSION["nombre"] = $_POST["nombre"];
    $_SESSION["apellido"] = $_POST["apelllido"];
    ...
?>

...
```

*Figura 14:* Los datos de un envío del formulario se asignan a dos variables de sesión

```
<?php
    session_start();
?>

...

<?php
    echo "¡", "Bienvenido", $_SESSION["nombre"], " ", $_SESSION["apellido"], "¡";
?>

...
```

*Figura 15:* Los datos previamente almacenados en una sesión se emiten

Al finalizar una sesión, PHP sólo tiene que llamar a la función `session_destroy()`; . Esto generalmente se realiza mediante una secuencia de comandos de cierre de sesión como la que se muestra en la figura 16. La última línea que se muestra en este script redirige al usuario a la página principal de la aplicación web.



```
<?php
    session_start();
    session_destroy();

    Header("Location: index.php");
?>
```

Figura 16: Script de cierre de sesión

Para asignar la sesión correcta cuando llega una solicitud a través de Internet se pueden utilizar diferentes métodos, aunque se utilizan con más frecuencia las denominadas *cookies* para identificar de forma exclusiva un navegador web. Una cookie es una pieza de información que se almacena en el equipo cliente que ejecuta el navegador web. Una aplicación web puede almacenar cualquier tipo de información sobre el navegador web. Sin embargo, debido a problemas de seguridad (es bastante sencillo falsificar una cookie) por lo general sólo un ID de sesión se almacena en el equipo del cliente. Los datos reales (nombre, apellido, etc.) se almacenan en el servidor. El ID de sesión se envía posteriormente junto con cada solicitud al servidor web, identificando de este modo el navegador web. En circunstancias normales, todo este proceso pasará desapercibido para el usuario. Sin embargo, es posible que surja el problema de que un usuario se queje de que la aplicación web no está funcionando correctamente. Esto puede ser debido a que el usuario tenga las cookies deshabilitadas previamente dentro del navegador web.

### 4.1.3 Bases de datos

El gestor de sesiones es un servicio muy útil para el almacenamiento de datos a corto plazo. Sin embargo, después de que el usuario se haya desconectado (o la sesión haya caducado) se pierde toda la información. Generalmente esto no supone un problema. Por ejemplo, si la tarea de la aplicación es el envío de un correo electrónico en respuesta a un envío de formularios HTML, no se necesita nada más allá de la gestión de sesiones. Sin embargo, en muchos otros casos es posible que desee almacenar información acerca de un usuario por más tiempo que sólo la duración de una sesión única, como por ejemplo en aplicaciones web que almacenan datos de los clientes, proporcionar un directorio de empleados, etc.

Como su nombre indica, las bases de datos almacenan datos de forma constante. Una base de datos es una parte del software que a menudo se ejecuta en el mismo equipo físico que el servidor web. En instalaciones más grandes de una base de datos también se puede ejecutar en una o más máquinas específicas para ello que se comunican con el servidor web.

El navegador web nunca se comunica directamente con una base de datos, sino solamente con el servidor web, que a su vez habla con la base de datos. Un script PHP envía o recibe

datos y crea una página HTML que contiene los resultados que se envían de vuelta al navegador web.

Existen muchas clases diferentes de bases de datos, solo las más relevantes para el proyecto actual se definen a continuación [27].

### 1. Bases de datos relacionales

Es el más común de todos los tipos diferentes de bases de datos. La manera de almacenar los datos en una base de datos relacional es mediante diversas tablas de datos. Cada tabla tiene un campo clave que se utiliza para conectarse a otras tablas. Por lo tanto, todas las tablas están relacionadas entre sí a través de varios campos clave.

Ejemplos de bases de datos relacionales son Oracle, Sybase y Microsoft SQL Server, y a menudo son partes clave del proceso de desarrollo de software.

### 2. Bases de datos operacionales

En su funcionamiento día a día, una organización genera una enorme cantidad de datos, tales como la gestión de inventario, compras, actas y finanzas. Todos estos datos se recogen en una base de datos que se puede conocer por distintos nombres, tales como base de datos operacional/ de producción, SADB (subject-area database) o bases de datos de transacción . Una base de datos operacional suele ser muy importante para las organizaciones, ya que incluye la base de datos de clientes, base de datos personal y base de datos de inventario. Los datos almacenados en bases de datos operacionales pueden ser modificados y manejados en función de lo que requiera la empresa.

Para realizar la página web en cuestión se necesitará una base de datos relacional, su función será almacenar la información de registro de cada usuario y empresa más la información personal de cada uno que le sea proporcionada, las conexiones de amistad entre usuarios y los miembros que son parte de una misma empresa. Una base de datos relacional posee las siguientes características [28]:

- Se compone de varias tablas o relaciones.
- No pueden existir dos tablas con el mismo nombre ni registro.
- Cada tabla es a su vez un conjunto de campos (columnas) y registros (filas).
- La relación entre una tabla padre y un hijo se lleva a cabo por medio de las claves primarias y claves foráneas (también denominadas ajenas).
- Las claves primarias son la clave principal de un registro dentro de una tabla y éstas deben cumplir con la integridad de datos.
- Las claves ajenas se colocan en la tabla hija, contienen el mismo valor que la clave primaria del registro padre; por medio de éstas se hacen las formas relacionales.

Hay muchos paquetes de software de bases de datos diferentes disponibles en la actualidad. Existe un tipo de software exclusivamente dedicado a tratar con bases de datos relacionales, conocido como Sistema de Gestión de Bases de Datos Relacionales (SGBDR, oRDBMS del inglés Relational Database Management System), también llamados manejadores o gestores de las BDR. A continuación analizaremos los más importantes y las características de cada uno:

### 1. Microsoft Access

Es quizás el más popular. Aunque Access es una base de datos de escritorio potente y fácil de usar, es poco adecuada para ser utilizada con aplicaciones web (por temas de rendimiento y acceso simultáneo principalmente) [28]. No se profundizará más en este tipo de base de datos, ya que no se ajusta a las necesidades de la página web que se quiere desarrollar.

### 2. MySQL

MySQL es una base de datos de código abierto ampliamente utilizado. MySQL está disponible de forma gratuita en virtud de varios acuerdos de licencia diferentes. Esta base de datos es altamente utilizada debido a su aplicación web, la cual está muy ligada a PHP, que a menudo aparece en combinación con MySQL. MySQL es una base de datos muy rápida en la lectura, pero puede provocar problemas de integridad en entornos de alta concurrencia en la modificación. En aplicaciones web hay baja concurrencia en la modificación de datos y en cambio el entorno es intensivo en lectura de datos, lo que hace a MySQL ideal para este tipo de aplicaciones. Sea cual sea el entorno en el que va a utilizar MySQL, es importante monitorizar de antemano el rendimiento para detectar y corregir errores tanto de SQL como de programación [29].

Algunas de las características de MySQL son las siguientes [30]:

- Amplio subconjunto del lenguaje SQL. Algunas extensiones son incluidas igualmente.
- Disponibilidad en gran cantidad de plataformas y sistemas.
- Posibilidad de selección de mecanismos de almacenamiento que ofrecen diferentes velocidades de operación, soporte físico, capacidad, distribución geográfica, transacciones...
- Transacciones y claves foráneas (limitaciones referenciales entre dos tablas).
- Conectividad segura.
- Replicación: Es posible copiar y mantener actualizados los datos en varios nodos.
- Búsqueda e indexación de campos de texto.

### 3. PostgreSQL

PostgreSQL es un Sistema de gestión de bases de datos relacional orientado a objetos y libre, publicado bajo la licencia PostgreSQL [31].

Las principales características de PostgreSQL son [32]:

- Alta concurrencia.
- Amplia variedad de tipos nativos.
- Claves foráneas.
- Disparadores (triggers): Un disparador o trigger se define como una acción específica que se realiza de acuerdo a un evento, cuando éste ocurra dentro de la base de datos. En PostgreSQL esto significa la ejecución de un procedimiento almacenado basado en una determinada acción sobre una tabla específica.
- Vistas.
- Integridad transaccional.
- Herencia de tablas.
- Tipos de datos y operaciones geométricas.
- Soporte para transacciones distribuidas. Permite a PostgreSQL integrarse en un sistema distribuido formado por varios recursos.

Debido a la facilidad que proporciona MySQL para entornos de desarrollo web y las características propias que se han enumerado previamente, se ha optado por utilizar este tipo de base de datos para la aplicación.

Muchos servicios de alojamiento web realizan una instalación de MySQL de trabajo a disposición de sus clientes, por lo tanto, a menudo se ahorra el esfuerzo de instalar el software de código de MySQL. Con el fin de administrar la base de datos (creación de las estructuras de datos, etc.), se puede utilizar la interfaz de usuario basada en la línea de comandos o herramientas de terceros más fácil de utilizar, tales como phpMyAdmin, de libre disposición y bastante popular. Como su nombre indica, phpMyAdmin es en sí misma una aplicación web basada en PHP y también se ofrece preconfigurado por muchos servicios de alojamiento web.

PhpMyAdmin es una herramienta de software libre escrita en PHP, creada con la intención de manejar la administración de MySQL a través de la Web. PhpMyAdmin es compatible con una amplia gama de operaciones de MySQL y MariaDB. Operaciones que se realizan frecuentemente (operaciones de gestión de bases de datos, tablas, columnas, relaciones, índices, usuarios, permisos, etc) se puede realizar a través de la interfaz de usuario, además de existir la posibilidad de ejecutar directamente cualquier sentencia SQL [33].

Las características de phpMyAdmin son las siguientes:

- Interfaz web intuitiva.
- Soporte para la mayoría de las características de MySQL:
  - Buscar y borrar bases de datos, tablas, vistas, campos e índices
  - Crear, copiar, eliminar, cambiar el nombre y modificar bases de datos, tablas, campos e índices

- Servidor de mantenimiento, bases de datos y tablas, con propuestas sobre la configuración del servidor
  - Ejecutar, editar y registrar cualquier sentencia SQL.
  - Administrar cuentas y privilegios de usuario de MySQL
  - Gestión de los procedimientos almacenados y disparadores
- Importar datos de CSV y SQL.
- Exportar datos a diversos formatos: CSV, SQL, XML, PDF, ISO / IEC 26300 - texto y hojas de cálculo de OpenDocument, Word, látex y otros.
- Administración de múltiples servidores.
- Creación de gráficos del diseño de la base de datos en varios formatos.
- Creación de consultas complejas utilizando búsqueda mediante ejemplo (QBE - Query by Example).
- Buscar globalmente en una base de datos o un subconjunto de ella.
- Transformar los datos almacenados a cualquier formato utilizando un conjunto de funciones predefinidas.

## 4.2 Ejecución del servidor

Un servidor es un programa de ordenador o un dispositivo que proporciona funcionalidad a otros programas o dispositivos, llamados clientes. Esta arquitectura se denomina como el modelo cliente-servidor, y un solo cálculo global se distribuye a través de múltiples procesos o dispositivos. Los servidores pueden proporcionar varias funcionalidades, a menudo llamadas servicios, tales como el intercambio de datos o recursos entre varios clientes, o realizar un cálculo para un cliente. Un solo servidor puede servir a varios clientes, y un solo cliente puede utilizar varios servidores. Un proceso de cliente puede ejecutarse en el mismo dispositivo o puede conectarse a través de una red a un servidor en un dispositivo diferente [34]. Algunos servidores típicos son los servidores de bases de datos, servidores de archivos, servidores de correo, servidores de impresión, servidores web, servidores de juego y servidores de aplicaciones [35].

Sistemas cliente-servidor son lo más frecuentemente implementado por el modelo de solicitud-respuesta: un cliente envía una solicitud al servidor, el cual realiza alguna acción y envía una respuesta de vuelta al cliente, por lo general con un resultado o contestación .

Para comenzar, se necesitará un servidor de bases de datos que proporcione la funcionalidad de mantener y compartir cualquier tipo de base de datos en una red [36]. Con este propósito y siendo el tipo de base de datos que se quiere utilizar MySQL, se ha decidido optar por XAMPP.

XAMPP es un servidor independiente de plataforma, software libre, que consiste principalmente en el sistema de gestión de bases de datos MySQL, el servidor web Apache y los intérpretes para lenguajes de script: PHP y Perl. También incluye otros módulos como OpenSSL y phpMyAdmin, del cual se habló en el punto anterior. El nombre proviene del acrónimo de X (que significa que sirve para cualquiera de los diferentes sistemas operativos), Apache, MySQL, PHP, Perl. Desde la versión 5.6.15, XAMPP cambió la base de datos de MySQL A MariaDB, pero la versión utilizada para este trabajo es la versión 3.2.2, en la cual se sigue utilizando MySQL.

El programa se distribuye bajo la licencia GNU y actúa como un servidor web libre, fácil de usar y capaz de interpretar páginas dinámicas.

De esta manera, se puede iniciar un servidor que funciona en el entorno local para realizar el desarrollo de la página web.

Después de la instalación de XAMPP se ha realizado la configuración del mismo y de los puertos a utilizar, y se ha iniciado el servidor, utilizando el panel de control que muestra la figura 17 a continuación.

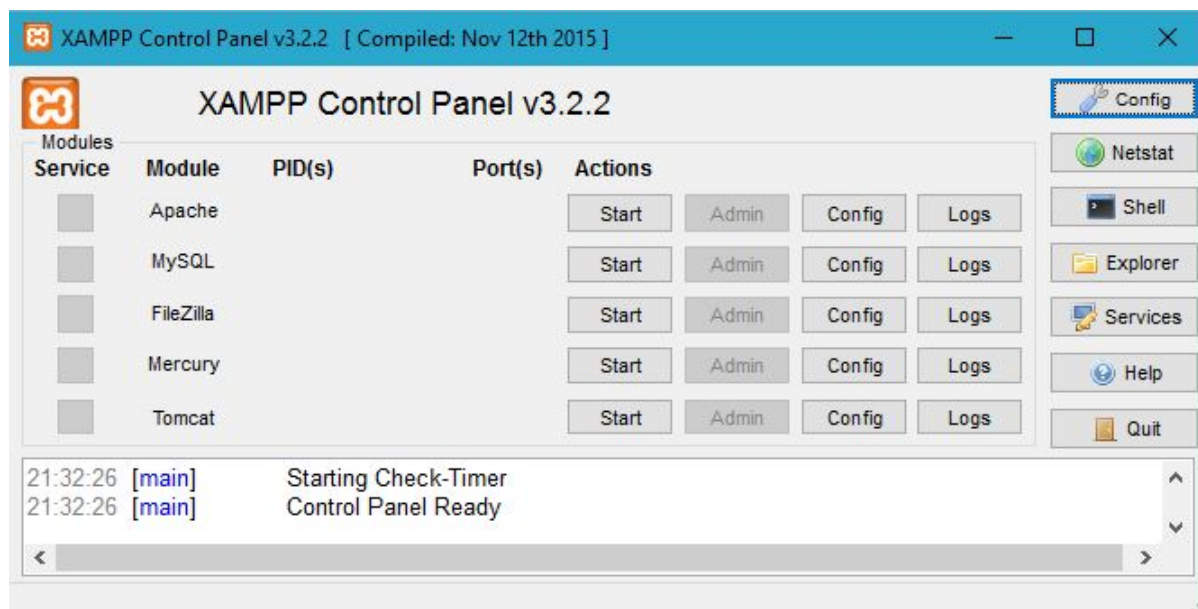


Figura 17: Panel de control de XAMPP

Al acceder a la dirección web <http://localhost:88/phpmyadmin> aparece la ventana que se observa en la figura 18, desde la cual se puede configurar la base de datos. Accediendo a este panel se pueden crear nuevas tablas, agregar elementos, modificar la información de las casillas...

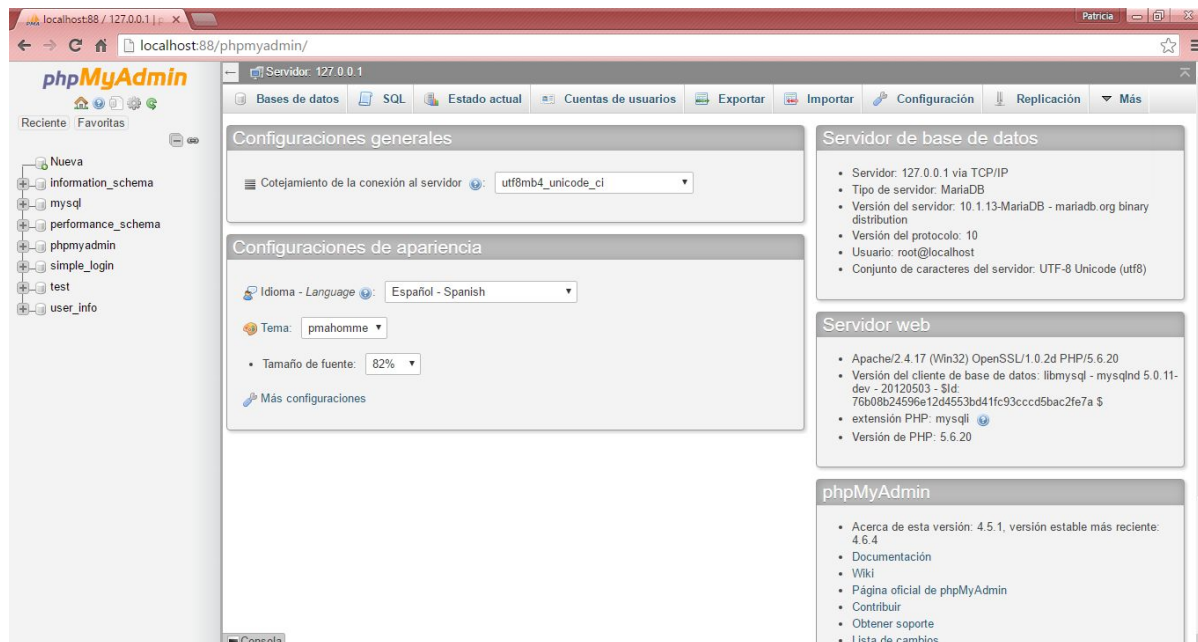


Figura 18: Pantalla de configuración de phpMyAdmin

De esta manera, se poseen los elementos necesarios para comenzar el desarrollo de la página web, tanto la programación de la parte front-end como el sistema que sostiene el back-end de la página. Para poder realizar este trabajo, se ha realizado una unión de la carpeta que contiene la información del framework Foundation y la carpeta de XAMPP, haciendo posible

de esta visualizar las páginas web que se programen con HTML, CSS, Javascript y PHP en el servidor local que se ha creado.



# Capítulo 5

## Creación de un buscador eficiente

En el presente capítulo se expondrá cómo se ha creado el elemento principal necesario en la página web: Un buscador que recoja la información de diferentes proveedores de MOOC y muestre al usuario la información relevante al introducir una palabra clave.

### 5.1 Investigación sobre creación de un buscador

La primera aproximación realizada por la autora consiste en la creación de un buscador web utilizando el lenguaje de programación php, pero una investigación sobre este tema mostró que para realizar el desarrollo de una función de búsqueda de este tipo era necesario contener la información a buscar en una base de datos, lo cual no se corresponde a las necesidades de la página web que se quiere crear, si no que los datos deben proceder de proveedores de MOOCs online.

Para esta funcionalidad, se ha determinado necesario utilizar una araña web o rastreador web. Una araña web es un bot, que busca sistemáticamente en Internet, por lo general con el propósito de indexación de la web (rastreo de la web).

Los motores de búsqueda web y algunos otros sitios utilizan una araña web o software de rastreo para actualizar su contenido web o índices de contenido web de otros sitios web. Las arañas web pueden copiar todas las páginas que visitan para un posterior procesamiento por un motor de búsqueda que indexa las páginas descargadas, por lo que los usuarios pueden buscar de manera mucho más eficiente.

#### 5.1.1 Construir una araña web

Hay bastantes factores a tener en cuenta en la construcción de un rastreador web, en general es un proceso fácil pero pueden surgir problemas cuando se quiere escalar el sistema, es decir, hacer que el sistema sea capaz de manejar el crecimiento continuo de trabajo de manera fluida [37] .

Ganlio detalla cómo sería una solución básica para construir una araña web rudimentaria, la cual se ejecutaría en una sola máquina con un solo hilo. Esta solución podría ser optimizada más adelante [38].

Para rastrear una sola página web, todo lo que se necesita es una petición GET HTTP a la URL correspondiente y analizar los datos de respuesta, los cuales son la esencia del rastreador. Con esto en mente, un rastreador web básico puede funcionar así:

- Comenzar con un conjunto de URL que contiene todos los sitios web que desea rastrear.
- Para cada URL, emitir una solicitud HTTP GET a buscar el contenido de la página web.
- Analizar el contenido (generalmente HTML) y extraer las direcciones URL potenciales que va a rastrear.
- Añadir nuevas direcciones URL al conjunto y continuar el rastreo.

Como se ha comentado anteriormente, cualquier sistema se enfrentará a un montón de problemas después de escalar. En una araña web, hay una gran cantidad de cosas que pueden salir mal cuando se escala el sistema a múltiples máquinas.

La frecuencia con la que se va a rastrear un sistema puede, en principio, no parecer un gran problema a menos que el sistema necesite nuevos contenidos en poco tiempo. Por ejemplo, si desea obtener los últimos cursos MOOC que se han anunciado, el rastreador puede necesitar mantener el rastreo de la página web del proveedor de cursos online cada hora. Esto podría suponer un problema para algunos sitios web pequeños, ya que es muy probable que sus servidores no puedan manejar dicha solicitud con tanta frecuencia. Un método para solucionar este problema es en seguir el robot.txt de cada sitio. Un robot.txt consiste básicamente en un estándar utilizado por los sitios web para comunicarse con los rastreadores web. Se pueden especificar tareas tales como qué archivos no deben ser rastreados, y la mayoría de los rastreadores web seguirán la configuración que establece [39]. Además, se pueden fijar diferentes frecuencias de rastreo para distintos sitios web. Por lo general, sólo unos pocos sitios necesitan ser rastreados varias veces al día.

En una sola máquina, se puede mantener el conjunto de URL en la memoria y eliminar entradas duplicadas. Sin embargo, en un sistema distribuido puede ser más complicado. Múltiples rastreadores pueden extraer la misma URL desde distintas páginas web y es posible que todos intenten añadir esta URL al conjunto de URL. Por supuesto, no tiene sentido rastrear la misma página varias veces, por lo que sería necesario realizar una deduplicación de datos. La deduplicación de datos se trata de una técnica especializada de compresión de datos que sirve para eliminar copias duplicadas de datos repetidos [40].

Un enfoque común es utilizar un filtro de Bloom. El filtro de Bloom se trata de una estructura de datos probabilística que nos permite saber si un elemento está contenido en un conjunto con un mínimo consumo de memoria. Sin embargo, puede resultar en falsos positivos. Por tanto, un filtro de Bloom puede decir si una URL definitivamente no está en el conjunto o si probablemente se encuentra dentro del conjunto de URL [41].

Después de recoger los datos de respuesta desde la página web, el siguiente paso es analizar estos datos (generalmente se trata de HTML) para extraer la información que se precisa. Esta parte puede resultar complicada si se desea que el sistema sea robusto.

El problema que surge es que se encontrarán lenguajes de marcado (markups) extraños, URLs, etc. en el código HTML y es difícil cubrir todos los casos extremos. Es posible que se

tengan que manejar problemas de codificación / decodificación cuando el HTML contiene caracteres que no son Unicode. Además, cuando la página web contiene imágenes, vídeos o incluso PDF, también es posible que se provoquen comportamientos extraños.

Asimismo, algunas páginas web están renderizadas a través de Javascript, como las que utilizan AngularJS, y es posible que el rastreador no sea capaz de obtener ningún contenido.

Para realizar este trabajo, existen servicios de rastreadores web de código abierto que permiten al usuario crear un sistema de búsqueda efectivo y fácil de desarrollar.

Los requisitos que la autora ha considerado necesarios o deseables para elegir un servicio de rastreador web son los siguientes:

- Que sea de licencia de uso libre
- Que tenga amplia documentación
- Que posea una interfaz gráfica de usuario o GUI
- Que sea capaz de funcionar en Windows

A continuación se desarrollan los servicios que son conocidos como los más importantes y populares.

### **5.1.2 Openwebspider**

Openwebspider [42] se trata de un rastreador web y motor de búsqueda de código abierto. Lo primero que se encuentra al acceder a la dirección web es una página sencilla con las últimas noticias, tal y como se puede observar en la figura 19.

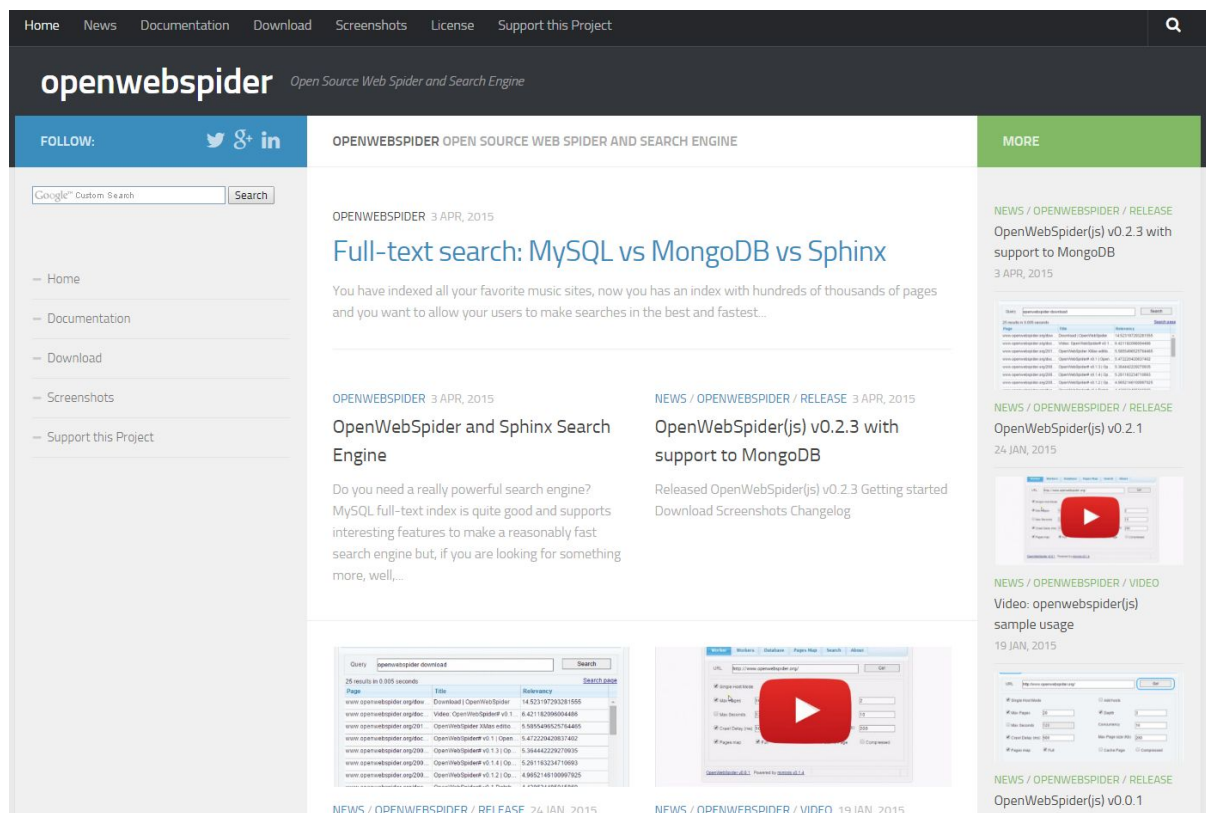
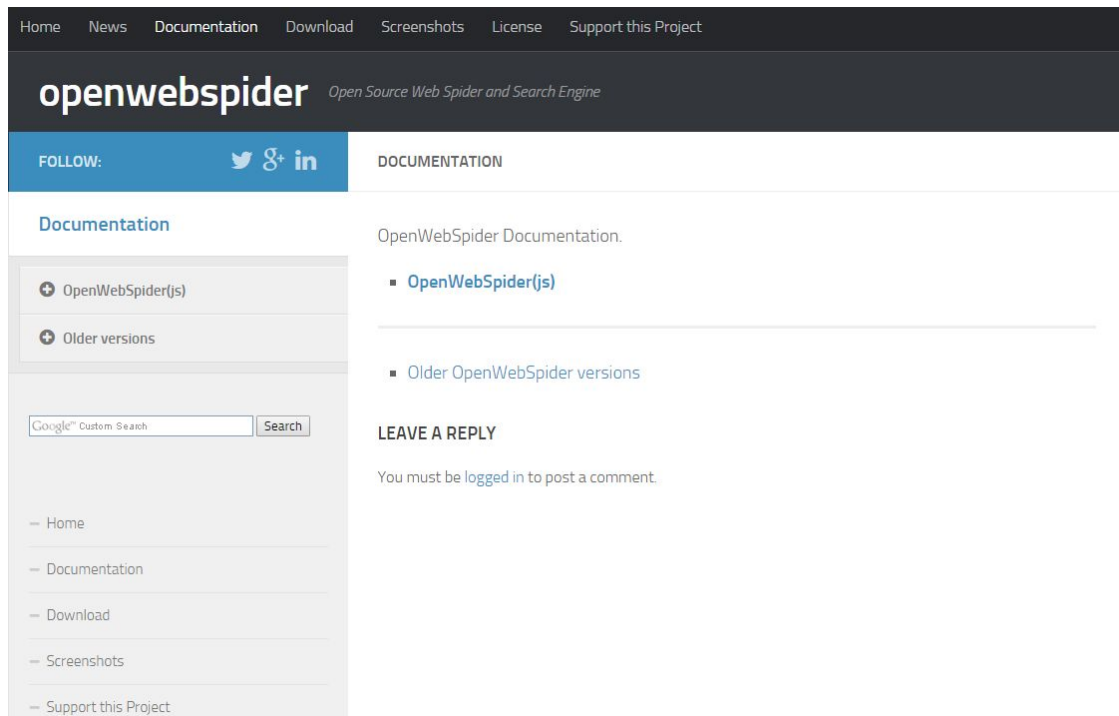


Figura 19: Openwebspider - Vista principal

Al acceder a la pestaña de documentación se puede ver un enlace que dirige a la información de la documentación disponible del servicio (figura 20). Esta documentación se trata de unos sencillos pasos para la ejecución del rastreador y comenzar su trabajo.



*Figura 20: Openwebspider - Documentación*

Según se observa al entrar en dicha documentación, al utilizar este rastreador web sólo se puede introducir una página web de donde sacar la información (figura 21), por tanto no sería adecuado para este trabajo, ya que se quiere recoger la información de diferentes proveedores MOOC.

Por otra parte, no se ha podido encontrar más documentación sobre el funcionamiento de esta araña web que la poca ofrecida en la web oficial.

3. Go in the third tab (Database) and configure your MySQL user, password, db, ...
4. Verify that openwebspider correctly connects to your MySQL server by clicking the "Verify" button
5. "Save" your configuration
6. "Create DB"; this will create all tables needed by openwebspider  
(remember that this will remove all existing tables and will create them from scratch)



7. Now you are ready to start an openwebspider worker; first tab (Worker): Go



*Figura 21: Openwebspider - Modo de empleo*

### 5.1.3 Bobik

Bobik [43] se trata de un servicio web que sirve para recopilar información de forma automática de la Web. Bobik emplea potentes mecanismos para realizar el trabajo en paralelo, es compatible con sitios web dinámicos (es decir, incluye los sitios generados a través de Ajax), y permite interactuar con él utilizando una API basada en REST.

El envío de una solicitud para rastrear a Bobik quiere decir que se va a hacer sólo una petición HTTP (a Bobik) y no se van a descargar los datos que no se van a utilizar (es decir, el 90% de la página).

Por tanto, se puede interactuar con Bobik en cualquier idioma y sin preocuparse de la cantidad de CPU / memoria / red local se requiere, ya que todo el trabajo de cálculo se llevará a cabo por Bobik [44].

Al acceder a la página web de Bobik se encuentra una interfaz sencilla (figura 22), con un botón desde el cual se puede iniciar sesión como usuario e información sobre el rastreador y la manera de utilizarlo.

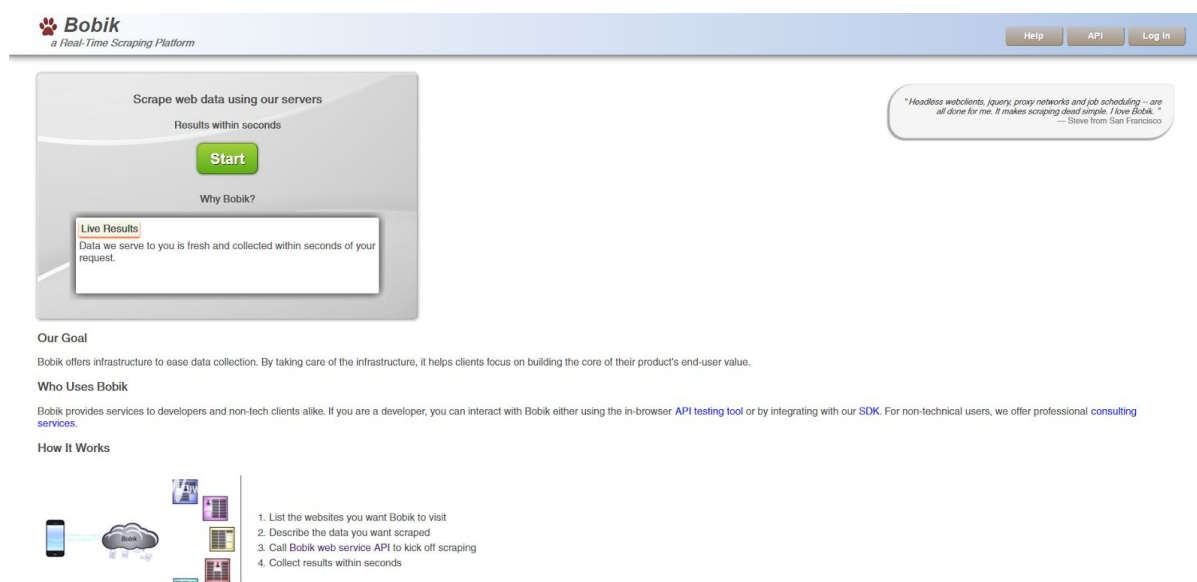


Figura 22: Bobik - Página de inicio

Bobik funciona de la siguiente manera:

- Primero se deben listar los sitios web que se desea que Bobik visite
- Describir los datos que desea rastrear
- Llamar a la API del servicio web de Bobik para comenzar el rastreo
- Recoger los resultados en cuestión de segundos

Bobik ofrece una amplia documentación como se observa en la figura 23, se puede acceder a ejemplos de su uso, un blog donde se encuentran ejemplos más avanzados y tutoriales, un

sistema de debug para poder realizar pruebas de la aplicación y ejemplos de cómo utilizar XPath.

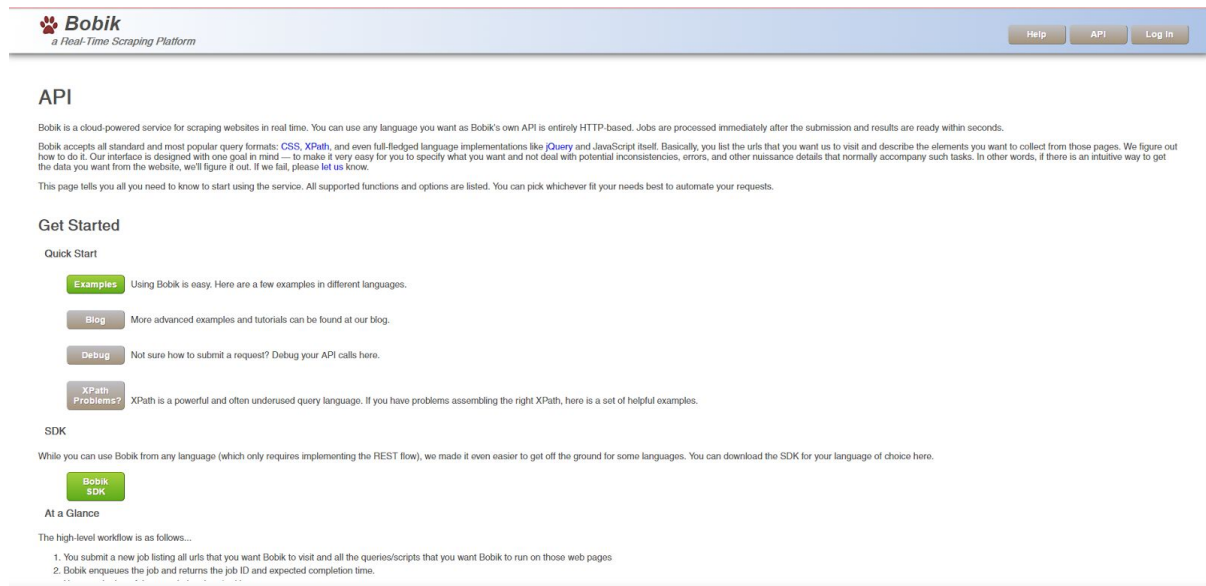


Figura 23: Bobik - Documentación

A primera vista, Bobik parece adecuada a las necesidades de la página de búsqueda de MOOCs. Sin embargo, al crear una cuenta e iniciar sesión se encuentra que la versión gratuita de Bobik sólo incluye un número de URLs por mes (figura 24) y las opciones de búsqueda que ofrece son escasas, no siendo posible por tanto adaptar los requisitos del buscador que se quiere desarrollar.

## Available Plans

200 urls Explorer	1,000 urls Basic	5,000 urls Standard	10,000 urls Power
<ul style="list-style-type: none"><li>Any number of urls per job</li><li>Any number of queries per job</li></ul>	<ul style="list-style-type: none"><li>Any number of urls per job</li><li>Any number of queries per job</li><li>Each additional url charged 1¢/url</li></ul>	<ul style="list-style-type: none"><li>Any number of urls per job</li><li>Any number of queries per job</li><li>Each additional url charged 1¢/url</li></ul>	<ul style="list-style-type: none"><li>Any number of urls per job</li><li>Any number of queries per job</li><li>Each additional url charged 1¢/url</li></ul>
Free	\$13 / month	\$45 / month	\$79 / month
<a href="#">Subscribed</a>	<a href="#">Subscribe</a>	<a href="#">Subscribe</a>	<a href="#">Subscribe</a>

## Billing Agreement

- Each plan covers a fixed number of urls at a flat monthly fee. Everything over is charged at a per-url rate.
- Every url, even the same one, counts as 1. That is because Bobik delivers you fresh data every time you use Bobik.
- The new subscription rate will be charged to the credit card on file immediately after you subscribe to a new plan.
- Bobik will continue charging this rate every month automatically unless the subscription is changed or cancelled.
- If you change the plan, unused amount from the old plan will be refunded to you as a courtesy.
- If you decide to cancel a subscription but miss the billing date, we'll refund you the money provided that no jobs were submitted in the interim.

Figura 24: Bobik - Planes disponibles



## 5.1.4 OpenSearchServer

OpenSearchServer [45] es un servidor de aplicaciones de código abierto que permite el desarrollo de aplicaciones basadas en índices tales como motores de búsqueda [46].

Las principales características que incluye OpenSearchServer son: Rastreador integrado para bases de datos, páginas web y documentos de texto enriquecido, una interfaz gráfica de usuario fácil de usar que permite el desarrollo de la mayoría de las aplicaciones a través de una interfaz de página Web construida en Zkoss, snippets, búsqueda por facetas, un procesador de HTML para la integración de los resultados de búsqueda en una página y funciones de supervisión y administración.

OpenSearchServer está escrito en Java y se puede integrar en casi cualquier tipo de aplicación sin necesidad de escribir código Java. Las APIs de REST / XML permiten que OpenSearchServer conecte con otros lenguajes de programación. La capacidad de aceptar plugins avanzados permite personalizaciones sofisticadas.

Al acceder a la web de OpenSearchServer se encuentran el acceso a la API de búsqueda y al software del motor de búsqueda y las características de cada uno, como se observa en la figura 25.

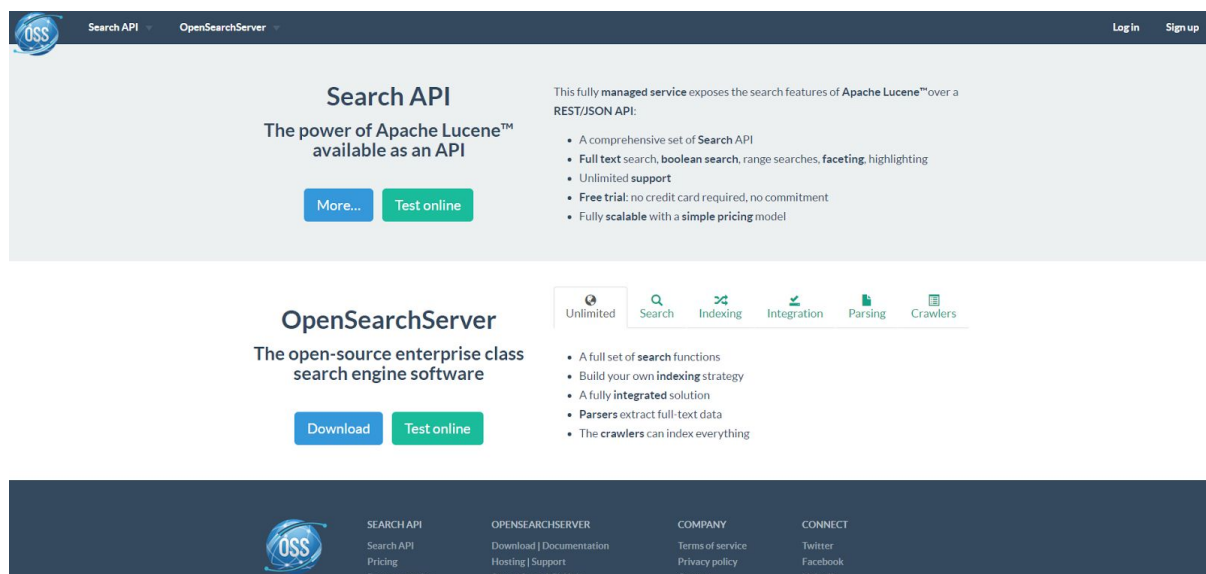


Figura 25: OpenSearchServer - Inicio

El acceso a la interfaz gráfica de usuario es de uso gratuito, ofreciendo la posibilidad de alojar una instancia en la infraestructura de OpenSearchServer mediante un pago mensual. Esta característica no será necesaria para el desarrollo del trabajo, ya que se va a alojar el servidor de manera local, tal y como se hizo en el caso del servidor para la base de datos (4.2).

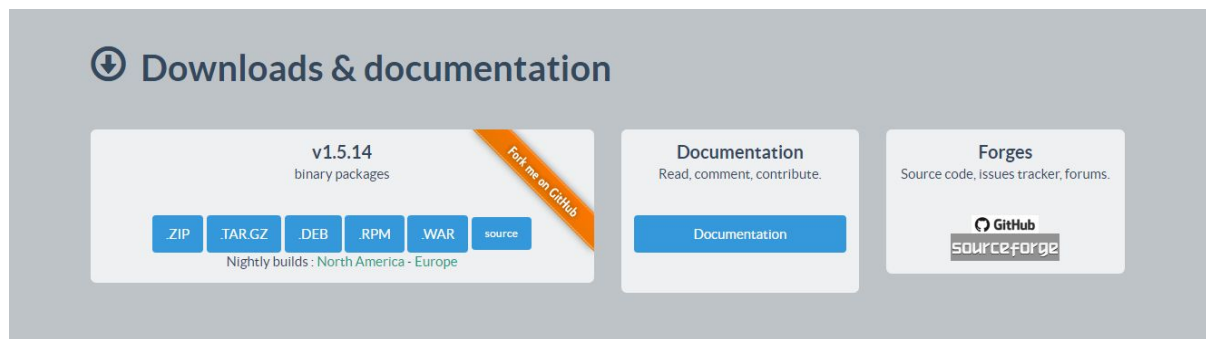


Figura 26: OpenSearchServer - Descargas y documentación

En la pestaña de documentación se pueden encontrar diferentes apartados (figura 27):

- Guía de instalación
- Compilar y generar
- Primeros pasos
- Preguntas frecuentes
- Bibliotecas de clientes
- Módulos y plugins
- API v1 (XML / HTTP)
- API v2 (RESTFul JSON)

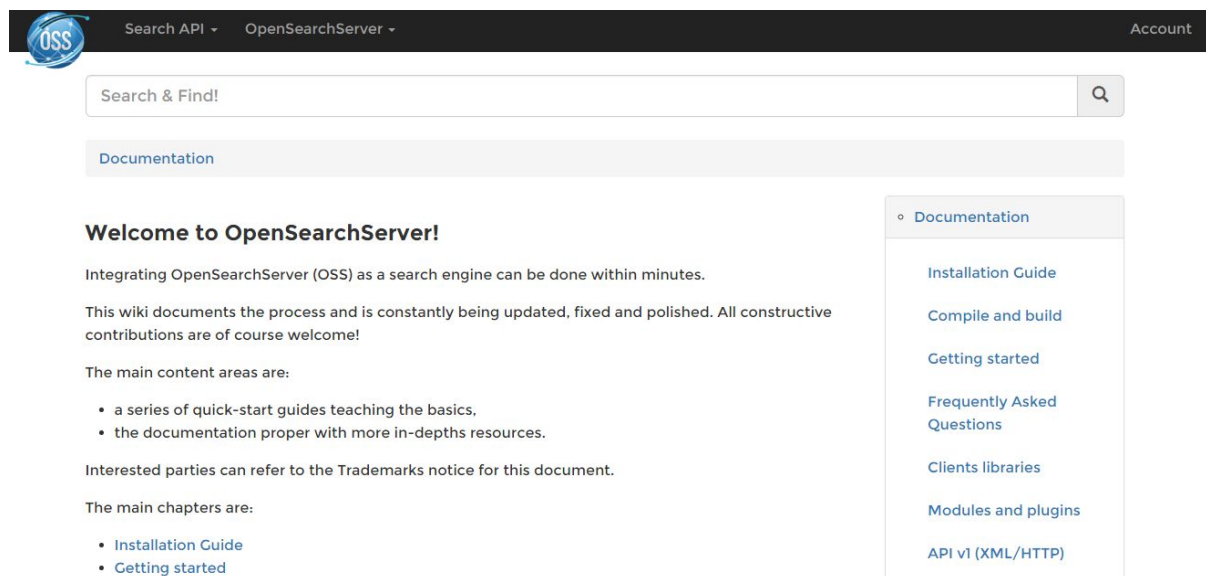


Figura 27: OpenSearchServer - Documentación

Las características observadas en OpenSearchServer cumplen con los requisitos buscados para el desarrollo del rastreador web que se quiere utilizar para el buscador, ya que se trata de una aplicación de código abierto gratuita que contiene una amplia documentación, una

interfaz de usuario muy completa y la posibilidad de instalación en Windows. Por tanto, se utilizará OSS para llevar a cabo el presente trabajo.

## 5.2 Desarrollo del motor de búsqueda

En el punto anterior se determinó el servicio a utilizar para realizar el rastreador web, OpenSearchService. Accediendo a la documentación [47] una vez iniciada sesión en la aplicación se pueden conocer los aspectos principales con los que cuenta y la manera de utilizarlo.

### 5.2.1 Conceptos principales

OpenSearchServer es un motor de búsqueda que se ejecuta en un servidor Windows, Linux o Solaris. Su interfaz gráfica de usuario se puede utilizar a través de cualquier navegador con soporte para Ajax (Internet Explorer, Firefox, Safari, Chrome). Dicha interfaz permite el acceso a todas las funciones de software libre. OSS también ofrece un conjunto completo de APIs REST y SOAP, facilitando la integración con otras aplicaciones. Las librerías cliente de PHP, Perl y ASP.NET permiten una fácil integración con entornos basados en PHP y Microsoft. OpenSearchServer ofrece además un módulo de Drupal y un plugin para Wordpress, y se puede integrar con estos sistemas CMS sin el trabajo de desarrollo.

### 5.2.2 Indexación

Para indexar el contenido, OpenSearchServer presenta los siguientes servicios:

- Rastreadores que busquen los datos según las normas que se les ha dado.
- Parseadores que extraen los datos para ser indexados (el texto completo) a partir de lo que se ha rastreado.
- Analizadores que aplican reglas semánticas y lingüísticas a la información indexada.
- Clasificadores que añaden información externa a los documentos indexados.
- Máquinas de aprendizaje automático que procesan los documentos indexados para deducir sus categorías.

#### Rastreadores

OpenSearchServer ofrece cuatro tipos de rastreadores:

- Rastreadores web.
- Rastreadores del sistema de archivos.
- Rastreadores de base de datos.
- Rastreador de archivos XML.

En principio se utilizará únicamente el rastreador web, ya que la información que se quiere rastrear proviene de diferentes agregadores MOOC online.

## Rastreador web

Este rastreador maneja HTTP y HTTPS, y por lo tanto se utiliza para rastrear sitios de Internet, intranet y extranet. Ofrece las siguientes funciones:

- Listado de sitios web. Una lista de direcciones URL, que puede incluir comodines (señalado con un carácter \*). Esto permite la indexación total o parcial de un sitio web. Se han utilizado para este caso dos direcciones web, como se puede ver en la figura 29: *[https://www.edx.org/course/\\*](https://www.edx.org/course/*)* y *[https://es.coursera.org/specializations/\\*](https://es.coursera.org/specializations/*)*, para poder abarcar todos los cursos que provengan tanto de edx como de coursera.
- Lista de exclusión. Otra lista de direcciones URL que pueden utilizar comodines, pero éste excluye páginas de la indexación. Las direcciones excluidas han sido *<https://www.edx.org/course/subject/>* y *[https://www.edx.org/course/\\*subject\\*](https://www.edx.org/course/*subject*)*, para que al buscar cursos en edx el buscador pueda filtrar solo los cursos, excluyendo las entradas de asignaturas.
- Filtrado de parámetros de las direcciones URL. Es posible ignorar datos sin importancia en las URL (por ejemplo, los parámetros de sesión).
- Navegador de direcciones URL. Esta interfaz se utiliza para revisar todas las direcciones URL conocidas y su estado (indexada, no indexada, error 404, etc.). En la figura 30 se observa la lista de enlaces relacionados con edx que se han visitado.
- Gestor de proceso de rastreo. Esta interfaz se utiliza para supervisar el proceso de indexación: número de hilos simultáneos, rastreo de velocidad para cada sitio, número de páginas por sesión, etc.
- Rastreo manual. Se utiliza para probar la indexación de una determinada URL.
- Proxy. Un proxy es un servidor, programa o dispositivo, que hace de intermediario en las peticiones de recursos que realiza un cliente a otro servidor [48]. Un proxy HTTP y/o HTTPS puede ser utilizado durante la indexación.
- Captura de pantalla. Toma una instantánea de las páginas indexadas.
- Autenticación. Para acceder a sitios web seguros.
- Protocolo de robots.txt. OpenSearchServer obedece las instrucciones de robots.txt, los cuales se mencionaron en el punto 5.1.
- Extracción de datos específicos. Tipos MIME (extensiones multipropósito de correo en Internet) [49] especificados, URLs, etc.

Index: web\_crawler Schema Query Renderer Update Delete Crawler Scheduler Runtime Reports Replication Privileges Inline help Refresh

Web Database Files REST

Pattern list Exclusion list URL filter URL browser Hostnames Crawl process Manual crawl Proxy Screenshot Authentication Field mapping

### Crawling parameters

User-Agent:  Number of URLs to crawl:

Fetch interval between re-fetches:   Maximum number of URLs per host:

Number of simultaneous threads:  Delay between each successive access, in seconds:

Job run when each session ends:

### Current status

Figura 28: OpenSearchServer - Interfaz gráfica [2.2]

Index: site Schema Query Renderer Update Delete Crawler Scheduler Runtime Reports Replication Scripts Privileges Inline help Refresh

Web Database Files Mailboxes REST

Pattern list Exclusion list URL filter URL browser Hostnames Crawl process Manual crawl Proxy Screenshot Authentication Field mapping Cookies Headers Scripts Robots.txt Site Map

☒ Enabled

For example:  
 Enter <http://www.opensearchserver.com> if you only want to crawl the home page  
 Enter [http://www.opensearchserver.com/\\*](http://www.opensearchserver.com/*) if you want to crawl all the content  
 Enter [http://www.opensearchserver.com/\\*support](http://www.opensearchserver.com/*support) if you only wish to crawl URLs containing the word "support" within the opensearchserver.com domain

Add

☒ [https://es.coursera.org/specializations\\*](https://es.coursera.org/specializations)

☒ [https://www.edx.org/course\\*](https://www.edx.org/course)

[ 1 - 2 / 2 ]

Figura 29: OpenSearchServer - Rastreador web [2.2]

URL(s) per page: 10 Total URL number: 21 Selected URL number: 21 Commands:   Buffer size: 10000

Url	Ev. Date	Mod. date	Type	Charset	Encoding	Content length	Lang	Lang method	Robots	Fetch	Code	Parsing	Index	Host	Links
<a href="https://www.edx.org/course">https://www.edx.org/course</a> *	16/09/16 13:03:24	16/09/16 12:38:18	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	1
<a href="https://www.edx.org/course/english-grammar-style-uq-wrte101x-2">https://www.edx.org/course/english-grammar-style-uq-wrte101x-2</a>	16/09/16 13:03:16	16/09/16 11:42:20	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2
<a href="https://www.edx.org/course/edeffi-test-preparation-insider-guide-etbi-itefi-0">https://www.edx.org/course/edeffi-test-preparation-insider-guide-etbi-itefi-0</a>	16/09/16 13:06:43	16/09/16 11:37:08	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	1
<a href="https://www.edx.org/course/programming-4-data-science-microsoft-dat209x-1">https://www.edx.org/course/programming-4-data-science-microsoft-dat209x-1</a>	16/09/16 13:06:44	16/09/16 11:43:40	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2
<a href="https://www.edx.org/course/analyzing-visualizing-data-excel-microsoft-dat208x-2">https://www.edx.org/course/analyzing-visualizing-data-excel-microsoft-dat208x-2</a>	16/09/16 13:03:54		text/html	UTF-8		-1	en	html lang attribute	Disabled	Permanent redirect	301	Parse	Indexed	www.edx.org	0
<a href="https://www.edx.org/course/introduction-programming-java-part-1-ucsmi-4-1x-1">https://www.edx.org/course/introduction-programming-java-part-1-ucsmi-4-1x-1</a>	16/09/16 13:03:55	16/09/16 12:00:41	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	1
<a href="https://www.edx.org/course/introduction-computer-science-mit-6-00-1x-8">https://www.edx.org/course/introduction-computer-science-mit-6-00-1x-8</a>	16/09/16 13:06:58	16/09/16 11:44:39	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2
<a href="https://www.edx.org/course/introduction-computer-science-harvard-cs50x">https://www.edx.org/course/introduction-computer-science-harvard-cs50x</a>	16/09/16 13:07:13	16/09/16 11:38:46	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2
<a href="https://www.edx.org/course/software-construction-java-mit-6-005-1x">https://www.edx.org/course/software-construction-java-mit-6-005-1x</a>	16/09/16 13:07:06	16/09/16 12:00:22	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2
<a href="https://www.edx.org/course/data-science-orientation-microsoft-dat101x-0">https://www.edx.org/course/data-science-orientation-microsoft-dat101x-0</a>	16/09/16 13:03:43	16/09/16 12:20:13	text/html	UTF-8		-1	en	html lang attribute	Disabled	Fetch	200	Parse	Indexed	www.edx.org	2

[ 1 - 10 / 21 ]

Figura 30: OpenSearchServer - Navegador de direcciones URL [2.2]

## Parseadores

El trabajo de un parseador o analizador sintáctico es la extracción de la información que va a ser indexada de los documentos que han obtenido los rastreadores. El analizador se selecciona automáticamente de acuerdo con el tipo MIME (si el rastreador puede encontrarlo)

o la extensión del archivo. A cada analizador se le puede dar un tamaño máximo para los archivos que se van a indexar.

Los siguientes formatos son admitidos por defecto:

- Archivos de Microsoft Word con extensión .doc (Word 6, Word 95, Word 97-2007) o .docx.
- Archivos de Microsoft Excel con extensión .xls (Excel 97-2007) o .xlsx.
- Archivos de Microsoft PowerPoint con extensión .ppt (PowerPoint 97-2007) o .pptx.
- HTML / XHTML.
- Archivos de OpenOffice con extensión .odt, .ods o .odf.
- Adobe PDF.
- Archivos RTF (Rich Text Format).
- Metadatos de archivos de audio con extensión .wav, .mp3, .aif o .OFF.
- Archivos .txt (la codificación se detecta de forma automática).

Más formatos pueden ser admitidos mediante los plugins de OpenSearchServer, desarrollados por la comunidad OSS.

Tales plugins incluyen:

- Archivos Torrent (indexación de sus metadatos).
- Archivos de Microsoft Publisher.
- Archivos de Microsoft VISIO.
- Archivos de Microsoft Outlook.

List of existing parser		
Audio	New parser	
Name	Type	Action
Audio parser	Audio	<a href="#">Edit</a> <a href="#">Delete</a>
DOC parser	DOC	<a href="#">Edit</a> <a href="#">Delete</a>
DOCX parser	DOCX	<a href="#">Edit</a> <a href="#">Delete</a>
FileSytm parser	File system	<a href="#">Edit</a> <a href="#">Delete</a>
HTML parser	HTML	<a href="#">Edit</a> <a href="#">Delete</a>
MSG parser	MAPI Msg (Outlook message)	<a href="#">Edit</a> <a href="#">Delete</a>
ODP parser		<a href="#">Edit</a> <a href="#">Delete</a>
ODS/ODF parser	ODS (OpenOffice spreadsheet)	<a href="#">Edit</a> <a href="#">Delete</a>
ODT parser	ODT (OpenOffice text file)	<a href="#">Edit</a> <a href="#">Delete</a>
PDF parser	PDF (Pdfbox)	<a href="#">Edit</a> <a href="#">Delete</a>
PPT parser	PPT	<a href="#">Edit</a> <a href="#">Delete</a>
PPTX parser	PPTX	<a href="#">Edit</a> <a href="#">Delete</a>
Publisher parser	PUB	<a href="#">Edit</a> <a href="#">Delete</a>
RSS parser	RSS	<a href="#">Edit</a> <a href="#">Delete</a>
RTF parser	RTF	<a href="#">Edit</a> <a href="#">Delete</a>

*Figura 31: OpenSearchServer - Parseadores existentes [2.2]*

Cada parseador proporciona información que se integra dentro del índice. Por ejemplo, los parseadores de Microsoft Office, OpenOffice y PDF incluyen específicamente un documento de título, el autor y el contenido. El analizador de HTML / XHTML identifica a cada enlace en la página, el título de cada página, etc.

Un sistema de mapeo de campo puede enviar estos datos a campos específicos, y la información de los parseadores puede ser procesada utilizando expresiones regulares. Por último, un campo puede presentar varios valores.



Fields Analyzers **Parser list** Stop words Synonyms Auto-completion Classifier Learner

Parser attributes Supported extension Supported MIME type URL pattern Field mapping

Name  Choose a name for this parser

Type HTML The type of this parser

Fail over  Choose a parser used in case of failure

Name	Value	
Size Limit	<input type="text" value="33554432"/>	The Size Limit of the file to be Parsed
Default Charset	<input type="text" value="UTF-8"/>	The charset to use if not charset is found
HTML parser	<input type="text" value="Best score"/>	The HTML parser to use
URL Fragment	<input type="text" value="remove"/>	The Option to remove or keep the URL Fragment
Ignore meta noindex	<input type="text" value="false"/>	Ignore the meta noindex tag in HTML header
Ignore meta nofollow	<input type="text" value="false"/>	Ignore the meta nofollow tag in HTML header
Ignore untitled documents	<input type="text" value="false"/>	Ignore documents which don't have title
Ignore non canonical pages	<input type="text" value="true"/>	Ignore documents which have a canonical link reference
Title Boost	<input type="text" value="2"/>	Boost for Title Tag in Html Content
H1 Boost	<input type="text" value="1.8"/>	Boost for H1 Tag in Html Content
H2 Boost	<input type="text" value="1.6"/>	Boost for H2 Tag in Html Content
H3 Boost	<input type="text" value="1.4"/>	Boost for H3 Tag in Html Content
H4 Boost	<input type="text" value="1.2"/>	Boost for H4 Tag in Html Content
H5 Boost	<input type="text" value="1.1"/>	Boost for H5 Tag in Html Content
H6 Boost	<input type="text" value="1.1"/>	Boost for H6 Tag in Html Content

Cancel Save

Figura 32: OpenSearchServer - Atributos de los parseadores [2.2]

Fields Analyzers **Parser list** Stop words Synonyms Auto-completion Classifier Learner

Parser attributes Supported extension Supported MIME type URL pattern **Field mapping**

linked in  captured by (reg.exp.):  ☐ remove tags Analyzer:

Parser field	Index field	Capture Reg.Exp.	Analyzer	Remove tags	Action
htmlSource	product_title	(?s)<div id="art_izq">[^\<]*<h1 class="clear">(.*?)</h1>		false	✖
htmlSource	product_description	(?s)<div class="clear"></div>[^\<]*<h5>[^\<]*</h5><p [^\<]*>(.*?)</p>		false	✖
htmlSource	product_image_url	(?s)(?.PRICE PRECIO)[^\<]*<b[^\<]*</b>[^\<]*<b>(.*?)&nbsp;		false	✖
htmlSource	product_reference	(?s)<h5>(?.Modelo Model):[^\<]*</h5>		false	✖

Cancel Save

Figura 33: OpenSearchServer - Mapeado de campo de los parseadores [2.2]

## Analizadores

Los analizadores procesan la información proporcionada por los parseadores. Su trabajo consiste en aplicar tratamientos semánticos y lingüísticos para que sea más fácil que las búsquedas encuentren resultados pertinentes. Por ejemplo, encontrar palabras ligeramente mal escritas. Se puede definir un número ilimitado de analizadores, y se aplican a los campos seleccionados dentro del índice.

Fields Analyzers Parser list Stop words Synonyms Auto-completion Classifier Learner

Choose an analyzer to edit

TextAnalyzer French Edit Delete

Editing analyzer: TextAnalyzer - FRENCH

Name: TextAnalyzer

Language: French

Tokenizer: LetterOrDigitTokenizerFactory

Additional characters

Filter	Scope	Properties	action
StandardFilter	Query and indexation		
LowerCaseFilter	Query and indexation		
ISOLatin1AccentFilter	Query and indexation		
FrenchStemFilter	Query and indexation		

Filter list:

Save Cancel

Analyzer test

Enter a text to analyze: Fernand ! lui cria-t-il, de mes cent noms, je n'aurais besoin de t'en dire qu'un seul pour te foudroyer ; mais ce nom, tu le devines, n'est-ce pas ? ou plutôt tu te le rappelles ? Query Test

Filter/Tokenizer	Result
LetterOrDigitTokenizerFactory	Fernand [1,8 - 1] lui [11,14 - 1] cria [15,19 - 1] t [20,21 - 1] il [22,24 - 1] de [26,28 - 1] mes [29,32 - 1] cent [33,37 - 1] noms [38,42 - 1] je [44,46 - 1] n [47,48 - 1] aurais [49,55 - 1] besoin [56,62 - 1]
LowerCaseFilter	fernand [1,8 - 1] lui [11,14 - 1] cria [15,19 - 1] t [20,21 - 1] il [22,24 - 1] de [26,28 - 1] mes [29,32 - 1] cent [33,37 - 1] noms [38,42 - 1] je [44,46 - 1] n [47,48 - 1] aurais [49,55 - 1] besoin [56,62 - 1]
ISOLatin1AccentFilter	fernand [1,8 - 1] lui [11,14 - 1] cria [15,19 - 1] t [20,21 - 1] il [22,24 - 1] de [26,28 - 1] mes [29,32 - 1] cent [33,37 - 1] noms [38,42 - 1] je [44,46 - 1] n [47,48 - 1] aurais [49,55 - 1] besoin [56,62 - 1]
FrenchStemFilter	fernand [1,8 - 1] lui [11,14 - 1] cri [15,19 - 1] t [20,21 - 1] il [22,24 - 1] de [26,28 - 1] me [29,32 - 1] cent [33,37 - 1] nom [38,42 - 1] je [44,46 - 1] n [47,48 - 1] aur [49,55 - 1] besoin [56,62 - 1]

Figura 34: OpenSearchServer - Editar un analizador [2.2]

Los analizadores incluyen un señalizador, el cual secciona los textos en tokens (cadena de caracteres que tiene un significado coherente en cierto lenguaje de programación [50]), y una serie de filtros configurables que procesan los textos tokenizados. Tales procesos incluyen:

#### Lematización

La lematización es un proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. Por ejemplo, "patinaje" también será indexado como "patín", aunque con una pertinencia menor.

OpenSearchServer puede lematizar palabras en 17 idiomas: Alemán, Inglés, árabe, chino, danés, español, finlandés, francés, holandés, húngaro, italiano, noruego, portugués, rumano, ruso, sueco y turco.

#### N-gramas

Los N-gramas parten las palabras en subsecuencias de una sílaba. Con n-gramas de tamaño dos, la palabra "patín" podría también ser indexado como "pa", "at", "ti", "in" y así sucesivamente. Esto se utiliza principalmente para identificar palabras ligeramente mal escritas.

#### N-gramas para palabras (Shingling)

Cuando los N-gramas definidos anteriormente son palabras enteras, esta técnica se denomina 'Shingling'. Consiste en identificar varias formas de decir una palabra que significan lo mismo. Por ejemplo, el acrónimo "I.B.M." o "I B M" o "I-B-M", etc. será indexado como "IBM".

#### Reducción de diacríticos

Esta técnica sustituye las letras que incluyen diacríticos y acentos con su equivalente base, sin alteraciones. Por ejemplo, la palabra francesa "côté" se convertirá en "cote".

#### Bloquear palabras

Este filtro hace caso omiso de ciertas palabras, que son consideradas de bajo valor en la información indexada. La lista de palabras bloqueadas se puede especificar para cada idioma.

#### Sinónimos

Este filtro localiza documentos mediante el uso de una lista parametrizable de sinónimos. Por ejemplo, la búsqueda de "júbilo" también podría proporcionar resultados como "alegría" o "jovialidad".

#### Mayúsculas y minúsculas

Este filtro ignora si las letras están en mayúsculas o minúsculas. Así, la palabra "Oso" también será indexada como "oso".

## Captura y sustitución con expresiones regulares

Este filtro aplica una expresión regular al texto. Un ejemplo típico es la transformación de una fecha con formato DD/MM/AAAA a una fecha AAAAMMDD, utilizando una expresión de captura `^([0-9] *)/([0-9]*)/([0-9] *)$` y siendo la cadena devuelta `$3$2$1`.

## Otros filtros

Existen muchos otros filtros disponibles, como la conversión grados a radianes, formateo de números, de-duplicación de palabras, extraer el nombre del dominio de una URL, normalizar direcciones URL, suministrar el título y la descripción de un vídeo de YouTube basado en su URL, etc.

## Clasificadores

Este módulo añade datos a un documento indexado. Se ejecutan una serie de consultas y cada consulta exitosa añade una palabra clave especificada para el documento en un campo determinado.

En la práctica, esto generalmente se utiliza para:

- Ordenar automáticamente los documentos indexados (asignándoles a un grupo de sitios, una categoría ...).
- Asociar palabras clave a los documentos.
- Gestionar los enlaces patrocinados.

The screenshot shows the 'Edit the selected classifier' window in OpenSearchServer. The classifier is named 'Category'. The 'Value is copied to field' is set to 'category'. The 'Score is copied to field' is empty. The 'Method' is set to 'Multivalued'. The 'Active' checkbox is checked. The 'Default value' is empty. Below these settings is a table with columns: Value, Boost, Query template, Query keywords, and Action. The table contains two rows of query templates.

Value	Boost	Query template	Query keywords	Action
Artisans Artisanat_d'art		hostname	www.example.org	
Artisans Artisanat_d'art		hostname	www.loremipsum.net	

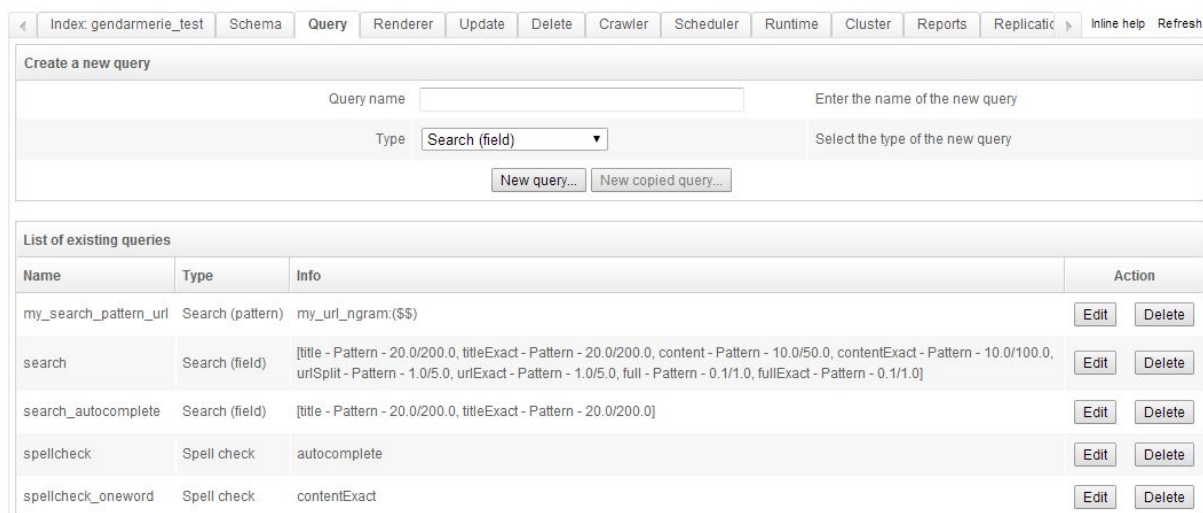
Figura 35: OpenSearchServer - Editar un clasificador [2.2]

## Máquinas de aprendizaje automático

Este módulo consta de dos partes. La primera aprende mediante el análisis de los documentos que ya han sido categorizados. Por ejemplo, una base de datos de artículos y categorías. La segunda parte se utiliza por lo general cuando se indexa un documento. La máquina analiza el documento, se basa en lo aprendido por el primer módulo, y sugiere una clasificación. Con ello, se puede por ejemplo sugerir categorías para un nuevo documento.

### 5.2.3 Consultas

El módulo de consulta puede crear cualquier número de plantillas de consulta, cada uno con sus propios parámetros.



The screenshot displays the OpenSearchServer Query management interface. At the top, there is a navigation bar with tabs: Index: gendarmerie\_test, Schema, Query (selected), Render, Update, Delete, Crawler, Scheduler, Runtime, Cluster, Reports, Replication, and Inline help. Below the navigation bar, the 'Create a new query' section contains a form with a 'Query name' input field, a 'Type' dropdown menu (set to 'Search (field)'), and two buttons: 'New query...' and 'New copied query...'. Below this, the 'List of existing queries' section shows a table with columns: Name, Type, Info, and Action. The table lists several queries, including 'my\_search\_pattern\_url', 'search', 'search\_autocomplete', 'spellcheck', and 'spellcheck\_oneword', each with its type and a set of 'Edit' and 'Delete' buttons.

Name	Type	Info	Action
my_search_pattern_url	Search (pattern)	my_url_ngram(\$\$)	<a href="#">Edit</a> <a href="#">Delete</a>
search	Search (field)	[title - Pattern - 20.0/200.0, titleExact - Pattern - 20.0/200.0, content - Pattern - 10.0/50.0, contentExact - Pattern - 10.0/100.0, urlSplit - Pattern - 1.0/5.0, urlExact - Pattern - 1.0/5.0, full - Pattern - 0.1/1.0, fullExact - Pattern - 0.1/1.0]	<a href="#">Edit</a> <a href="#">Delete</a>
search_autocomplete	Search (field)	[title - Pattern - 20.0/200.0, titleExact - Pattern - 20.0/200.0]	<a href="#">Edit</a> <a href="#">Delete</a>
spellcheck	Spell check	autocomplete	<a href="#">Edit</a> <a href="#">Delete</a>
spellcheck_oneword	Spell check	contentExact	<a href="#">Edit</a> <a href="#">Delete</a>

Figura 36: OpenSearchServer -Consultas [2.2]

## Tipos de consultas

Se pueden encontrar dos tipos diferentes de consultas:

### Consultas de campo

La búsqueda de consultas de campo simplifica la construcción de plantillas de consultas. Para utilizarlas, se ha de seleccionar los campos del esquema en el que se ejecutará la búsqueda de texto completo, y el peso relativo de cada uno de los campos cuando se calcula la relevancia de un documento.

Cada campo puede estar configurado de cuatro maneras:

- Por patrón: las palabras clave se mantienen sin transformar. Este método es parecido a las consultas patrón, que se definirán más adelante.
- Por palabras: los caracteres extraños en la consulta (comillas, símbolos, etc.) se eliminan, y la búsqueda se ejecuta en cada palabra de la consulta. Por ejemplo, una búsqueda de lorem ipsum sería en realidad una búsqueda de lorem seguido por una búsqueda de ipsum.
- Frase: la consulta se realiza como se ha visto en la anterior, pero se buscará la frase completa. Así, una búsqueda de lorem ipsum sólo buscará lorem ipsum.
- Palabra y Frase: Una combinación de los dos modos anteriores. Así, una búsqueda de lorem ipsum ejecutará tres sucesivas búsquedas para lorem, ipsum y finalmente lorem ipsum.

Field	Mode	Term boost	Phrase Boost
title	Pattern ▼	20	200
titleExact	Term & phrase ▼	20	200
titlePhonetic	Term & phrase ▼	0,1	0,1
content	Term & phrase ▼	10	50
contentExact	Term & phrase ▼	10	100
contentPhonetic	Term & phrase ▼	0,1	0,1
urlSplit	Term & phrase ▼	1	5
urlExact	Term & phrase ▼	1	5
urlPhonetic	Term & phrase ▼	1	5
full	Term & phrase ▼	0,1	1
fullExact	Term & phrase ▼	0,1	1
fullPhonetic	Term & phrase ▼	0,01	1

Figura 37: OpenSearchServer - Campos consultados [2.2]

## Consultas por patrón

La búsqueda de consultas por patrón hace uso de un lenguaje de consulta específico para designar los campos en los que será buscado el texto completo.

Estos son los elementos de sintaxis que se utilizan para calcular la relevancia:

- Operadores booleanos, es decir AND/OR/NOT + -. Por ejemplo, la consulta + Madden - 2012 busca documentos que debe contener la palabra Madden pero no la palabra de 2012.
- Búsqueda por proximidad. El uso de comillas en la consulta permite la búsqueda de una frase en lugar de las palabras por separado. Se puede especificar una tolerancia para la distancia entre estas palabras.
- Buscar en campos específicos. La búsqueda sólo cubrirá determinados campos, como el título de los documentos.
- Búsqueda ponderada. Un peso relativo se puede asignar a cada componente de la búsqueda, para dar más importancia a una palabra o un campo. Esto es comúnmente utilizado para asignar un peso mayor al título de un documento.
- Rango de búsqueda. Una búsqueda puede ser limitada entre dos fechas específicas, o, más en general, entre dos valores especificados.

Ejemplo de una plantilla de consulta:

```
title:($$)^10 OR title:("$$$")^10 OR  
url:($$)^5 OR url:("$$$")^5 OR  
content:($$) OR content:("$$$")
```

*Figura 38: Plantilla de consulta [2.2]*

En las consultas basadas en esta plantilla:

- Los strings \$\$ se sustituyen por las palabras clave que el usuario está buscando.
- Si estas palabras se encuentran en el campo de título o el campo URL, reciben un peso diez veces o cinco veces para los cálculos de relevancia.
- Las palabras clave se buscan como entidades independientes, pero también como una frase (búsqueda de proximidad utilizando comillas).

## Facetas y filtros

Estas dos funciones a menudo se utilizan conjuntamente, y por tanto serán presentadas al mismo tiempo.

## Facetas

Cuentan el número de resultados por entidad. Una "entidad" puede ser una categoría, el nombre de un dominio, un número determinado de enlaces, fechas, etc. Cualquier valor indexado dentro de un campo puede convertirse en una faceta.

Típicamente se utilizan para proporcionar el número de resultados por página web en la página de resultados de búsqueda, o el número de resultados por concepto, por una serie de intervalos de tiempo especificados, etc.

Las facetas pueden trabajar de manera individual o estar agrupadas, como se puede ver en el ejemplo de la figura 39.

The screenshot displays a search results interface for the term "moellon". At the top, a dark banner reads "RESULTATS DE RECHERCHE". Below this, a message states: "Votre recherche pour «moellon» sur tout le site a retourné les résultats suivants :".

**Affiner votre recherche**

- MAÇONNERIE GROS-OEUVRE (16)
  - Parpaings, briques, blocs béton divers (15)
    - Parpaings, bloc béton divers (15)
- Ciment, chaux, enduits poudres divers (1)
- OUTILLAGES & CONSOMMABLES (9)
- PLÂTRERIE, ISOLATION (1)

**VOS FILTRES**

SÉLECTION EN COURS :

MOT CLÉ : moellon

NOUVELLE RECHERCHE

Nouvelle recherche **ok**

**PRODUITS**  
26 résultats

26 résultats trouvés

Vos filtres : moellon

Trier par : pertinence 12 résultats par pages LISTE GALERIE

**37%** Parpaing chainage 20x20x50cm  
Réf 1310  
[> Voir la fiche produit](#)

Parpaing d'angle creux 20x20x50cm  
Réf 1319  
[> Voir la fiche produit](#)

Figura 39: Ejemplo del uso de facetas [2.2]



Field	Multivalued	Post collapsing	Minimal count	Action
product_parutionstate	no	no	2	remove
product_idsupport	no	no	2	remove
ouvrage_idlang	no	no	2	remove
ouvrage_maintheme	yes	no	20	remove
ouvrage_theme	yes	no	20	remove
product_idcover	no	no	2	remove
product_parutionyear	no	no	2	remove
product_instock	no	no	1	remove
firstcatalogprice	yes	no	2	remove
author_exact	yes	no	2	remove

product\_id add facet

Search

Figura 40: OpenSearchServer - Facetas [2.2]

## Filtros

Esta función normalmente va unida a las facetas. Una búsqueda se puede filtrar utilizando una subconsulta escrita en el mismo lenguaje de consulta utilizado en el patrón de consulta que se observó en la sección anterior.

Por ejemplo, una búsqueda podría estar limitada a un nombre de dominio, un intervalo de tiempo dado, etc.

#	Description	Negative	Action
1	Relative date filter	<input type="checkbox"/>	Add

Field	Date format	From interval	From unit	To interval	To unit
crawlDate	YYYYMMDD	5	days	1	days

Figura 41: OpenSearchServer - Filtros [2.2]

## Agrupaciones

Los documentos pueden ser agrupados para hacer la lista de los resultados más inteligible y menos repetitivo. Por ejemplo, se podría limitar a tres el número máximo de resultados consecutivos de un determinado sitio web.

Los diferentes tipos de agrupación son:

- Agrupación sucesiva optimizada: limita el número de resultados consecutivos en los resultados de búsqueda, optimizado para grandes volúmenes de resultados.
- Agrupación sucesiva completa: limita el número de resultados consecutivos a lo largo de todo el conjunto de resultados de búsqueda (y no sólo dentro de una determinada página de resultados).
- Agrupación simple: limita el número de resultados, ya sean consecutivos o no, a lo largo de todo el conjunto de resultados de búsqueda.

## Fragmentos de texto

Durante la renderización de los resultados de búsqueda, OpenSearchServer puede devolver el campo indexado tal y como es o crear un fragmento de texto inteligente.

Los fragmentos de texto siguen las siguientes pautas:

- Número máximo de caracteres. El fragmento no será más largo que el número elegido.
- Resaltar las palabras clave buscadas. La etiqueta utilizada para resaltar palabras en el listado de resultados es parametrizable.
- Relevancia del fragmento. OpenSearchServer devuelve la frase más relevante de las palabras clave buscadas, mientras que mantiene la longitud tope acordada.
- Detección de frase. OpenSearchServer intenta retroceder al comienzo de una frase relevante para utilizarla como fragmento.

Search (pattern): search

Query Patterns Returned fields Faceted fields **Snippet fields** Sorted fields Filters Geo Collapsing Boosting qu

Field	Fragmenter	Tag	Separator	Max snippet size	Max snippet number	Action
title	NoFragmenter	strong	...	200	1	Remove
author	NoFragmenter	strong	:	200	1	Remove
ouvrage_sumary	NoFragmenter	strong	...	20000	1	Remove
ouvrage_table	NoFragmenter	strong	...	20000	1	Remove
subtitle	NoFragmenter	strong	...	200	1	Remove
ouvrage_treatise	NoFragmenter	strong	...	200	1	Remove
ouvrage_collection	NoFragmenter	strong	...	200	1	Remove

product\_name ▼ add snippet

Search

Figura 42: OpenSearchServer - Fragmentos de texto [2.2]

## Clasificación

El orden por defecto de los resultados de un motor de búsqueda se realiza utilizando la puntuación de relevancia del algoritmo de búsqueda (Modelo de espacio vectorial). Como se mencionó anteriormente, este resultado puede ser refinado utilizando el lenguaje de consulta. La clasificación también se puede basar en un campo determinado, igual que se realiza en una base de datos. Un ejemplo consiste en ordenar los resultados en orden cronológico inverso. Las puntuaciones y los campos también se pueden mezclar. Por ejemplo, los resultados con una puntuación idéntica podrían ser ordenados en orden cronológico.

## Geolocalización

Los filtros de geolocalización pueden ser fácilmente integrados en una consulta, siempre que el esquema del índice incluya los campos de ubicación (latitud y longitud).

Con un filtro de este tipo, sólo se devolverán los documentos dentro de un rectángulo geográfica determinado. La distancia entre los documentos y el punto central también será calculada y devuelta.

Las coordenadas pueden ser expresadas en grados o en radianes, y las distancias en kilómetros o millas.

## Mejorar subconsultas

Mejorar las subconsultas ajusta la puntuación de relevancia de los documentos utilizando parámetros arbitrarios.

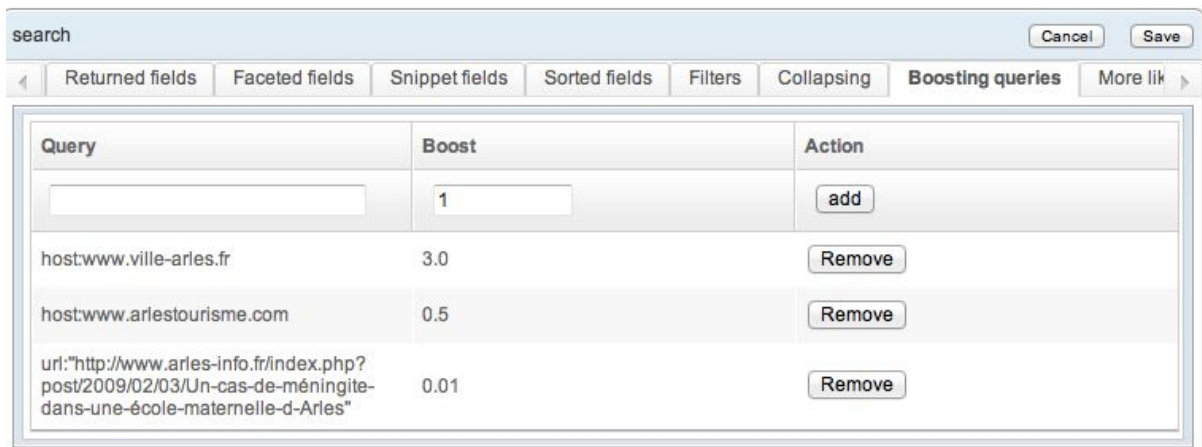


Figura 43: OpenSearchServer -Mejorar subconsultas [2.2]

Estas subconsultas pueden tanto reforzar o disminuir la puntuación de un documento. Los resultados de un determinado sitio web por lo tanto pueden ser favorecidos, o documentos de una categoría determinada pueden tener su puntuación disminuida. Múltiples consultas pueden aplicarse a cualquier búsqueda.

## Uniones

Una búsqueda puede unirse con uno o más índices para obtener más información. Estas uniones se pueden ajustar de forma precisa, y la información se pueden filtrar y ordenar antes de ser mostrada.

## Documentos relacionados

Esta función devuelve documentos similares a un documento dado. Este cálculo se basa en un corpus compartido de palabras y conceptos.

Este módulo ofrece distintas posibilidades de ajuste:

- Longitud máxima y mínima de las palabras dentro del corpus.
- Lista de palabras bloqueadas.
- Frecuencia más baja aceptable para una palabra dentro del corpus.

## Corrector ortográfico

El módulo de corrección de ortografía sugiere palabras alternativas. Los algoritmos disponibles son:

- Distancia de Jaro-Winkler.
- Distancia de levenshtein.
- Distancia de N-gramas.

## Extracción de entidades nombradas

Esta consulta puede extraer entidades nombradas de un índice relevante. Esta consulta puede a su vez ser utilizada dentro de otro índice, a través de un filtro analizador.

Esto permite que los sistemas automatizados puedan extraer entidades nombradas a partir de los textos indexados. Por ejemplo, un sistema podría tener dos índices, uno con entidades nombradas (como listas de nombres de ciudades o nombres de persona) y el otro con artículos de prensa. Para cada nuevo artículo que se está indexando, el segundo índice podría ejecutar una consulta de extracción de entidades nombradas en el primer índice, y grabar en un campo determinado del esquema una lista de ciudades y personas mencionadas en el artículo.

## Autocompletar

La función de autocompletar palabras se puede configurar en pocos pasos. Los campos utilizados para generar sugerencias se pueden elegir libremente.

OpenSearchServer cuenta con un motor de auto-completado pre-ensavado, que realiza el procesamiento de los datos indexados de una manera optimizada para ofrecer las propuestas más relevantes.

Se pueden configurar múltiples índices de autocompletado, por ejemplo para ofrecer a los usuarios de diferentes áreas sugerencias tomadas de diferentes campos dentro del mismo índice.

Name	Status	Terms	Command
autocomplete	STOPPED	7260	<button>Build</button>

Figura 44: OpenSearchServer - Autocompletado [2.2]

## 5.2.4 Renderización de la página de búsqueda

El módulo de Renderización de OpenSearchServer ofrece un rápido acceso a una página completa de resultados de búsqueda. Esta página se puede personalizar fácilmente utilizando CSS, como se muestra en la figura 46. La información que se debe mostrar para cada resultado se puede seleccionar con precisión. Se pueden incrustar en los resultados enlaces e imágenes.

Index: web\_crawler Schema Query **Renderer** Update Delete Crawler Scheduler Runtime F Inline help Refresh

Edit the renderer: default Fields CSS Style Log Authentication Viewer Testing

Edit the renderer: default

Renderer name: default Request name: search

Search button label: Search No result text: No result found

One result text: 1 result found Many results text: results found

Facet width: 200px Autocompletion: autocomplete

Header (HTML) Footer (HTML)

Cancel Save Save & close

Figura 45: OpenSearchServer - Renderizador [2.2]

Edit the renderer: default Fields Filters **CSS Style** Sorts Log Authentication Viewer Testing

```
text-align: center;
}

oss-facet {
padding: 1em;
}

ossfieldrd1 {
font-size: 120%;
}

ossfieldrd3 {
color: #0E774A;
}

oss-result {
padding: 25px;
background-color: #F5F5F5;
border-radius: 10px;
}

body {
background-color: #ff916;
}

container {
background-color: #ff916;
}

title {
margin: 30px;
margin-left: 170px;
text-align: center;
font-size: 24px;
color: #fff;
}

img {
display: none;
}
```

Restore default values

Cancel Save Save & close

Figura 46: OpenSearchServer - Estilos CSS del renderizador

## Autenticación

Un procedimiento de inicio de sesión se puede configurar para las páginas de resultados de búsqueda. Esto permite clasificar los resultados de búsqueda basado en perfiles de usuario (por nombre de usuario y / o por grupo), utilizando los valores almacenados en campos específicos del esquema durante la indexación.

El módulo de Renderización viene con esta opción incorporada, e incluye un formulario de autenticación.

## 5.2.5 Herramientas auxiliares

Las siguientes herramientas se utilizan para ejecutar procesos a gran escala. Permiten el control del motor de búsqueda, la automatización de tareas, y la generación de estadísticas sobre su uso.

### Automatización

El módulo de Planificación puede automatizar procesos complejos. Los usos más comunes son para implementar estrategias de indexación específicas, o copiar un mismo índice a través de múltiples servidores. Las tareas pueden ser programadas utilizando la sintaxis de crontab. El módulo de Planificación puede ejecutar las siguientes tareas:

- Lanzar un rastreador de bases de datos.
- Rastrear documentos XML.
- Iniciar / detener un rastreador web.
- Iniciar / detener un rastreador de un sistema de archivos.
- Optimizar la indexación.
- Reproducir la indexación.
- Consultas de borrado.
- Cargar registros.
- Eliminar todo el contenido de un índice.
- Descargar un archivo XML desde un servidor FTP para el procesamiento de XSL.
- Combinar índices.
- Ejecutar una tarea dentro de otro índice (local o remoto).
- Importar documentos desde otro índice (local o remoto) para ser reprocesados para la indexación.
- Re-indexar palabras dentro de los índices de autocompletación.

Figura 47: OpenSearchServer - Planificador [2.2]

## Hacer réplicas y crear backups

El módulo de Reproducción permite copiar un índice en el mismo servidor o en un servidor remoto. Esto se puede hacer manualmente o mediante el Planificador.

Los casos de uso típicos para este módulo son:

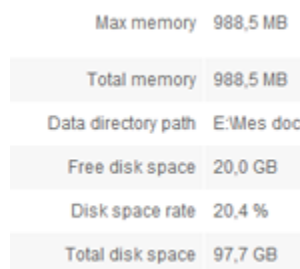
- Distribución de un índice sobre un grupo de servidores, sin interrumpir su funcionamiento.
- Creación de copias de seguridad de forma local o en un clúster remoto.

Figura 48: OpenSearchServer - Reproducción [2.2]



## Monitorización

El módulo de Tiempo de ejecución proporciona datos útiles, tales como el uso de memoria o espacio libre en el disco, y se integra dentro de la interfaz gráfica de usuario. Una API XML / HTTP está disponible para interactuar con otra aplicación de monitorización.



Max memory	988,5 MB
Total memory	988,5 MB
Data directory path	E:\Mes doc
Free disk space	20,0 GB
Disk space rate	20,4 %
Total disk space	97,7 GB

*Figura 49: OpenSearchServer - Monitorización [2.2]*

## Estadísticas

La herramienta de estadísticas incluye dos módulos: los registros diarios y la interfaz gráfica.

### Registros diarios

Los registros se encargan de indexar los siguientes datos para cada búsqueda:

- Marca de tiempo.
- Palabras clave.
- Número de documentos devueltos.

Cuando se combina con el módulo central, se dispone de la siguiente información adicional:

- ID de sesión.
- Dirección IP remota.
- Nombre de usuario.

Los archivos de registro se pueden exportar a aplicaciones de gestión de estadísticas.

### Interfaz gráfica

Esta interfaz muestra que palabras clave son las buscadas con más frecuencia y para que búsquedas no se produce ningún resultado.

La visualización puede ser limitada a un período de tiempo dado, y el informe se puede exportar como un archivo CSV.

# Capítulo 6

## Resultados finales

### 6.1 Página web final

Realizando una unión de toda la información explicada durante el trabajo se puede llevar a cabo el desarrollo de la página web.

Para ello, se ha utilizado el entorno XAMPP como base, tal y como se determina en el punto 4.2 y foundation como framework, como se especifica en el punto 3.1.4. Para realizar la web, se han seguido los diseños establecidos en el punto 2.3.2, dando lugar al resultado que se muestra a continuación.

Como nombre para la página se ha utilizado “CourseJam”, esto se trata de un nombre de prueba realizado por la autora con el objetivo de realizar las pruebas de la página web.

Para comenzar, se puede observar la pantalla final de inicio de sesión del programa en la figura 50, donde se dispone de distintos accesos según se trate de un empleado o una empresa. A continuación se encuentra el mockup previamente diseñado (figura 51), para comparación con el resultado final.

Para guardar las características de cada tipo de usuario se han creado dos tablas distintas en la base de datos (figura 52), ‘business’ guardará la información referente a empresas (figura 53), mientras que ‘member’ hará lo equivalente con la información de los usuarios (figura 54). Si se intenta iniciar sesión con datos que no existen en la base de datos, se muestra un cartel que informa de que el usuario y la contraseña no existen.

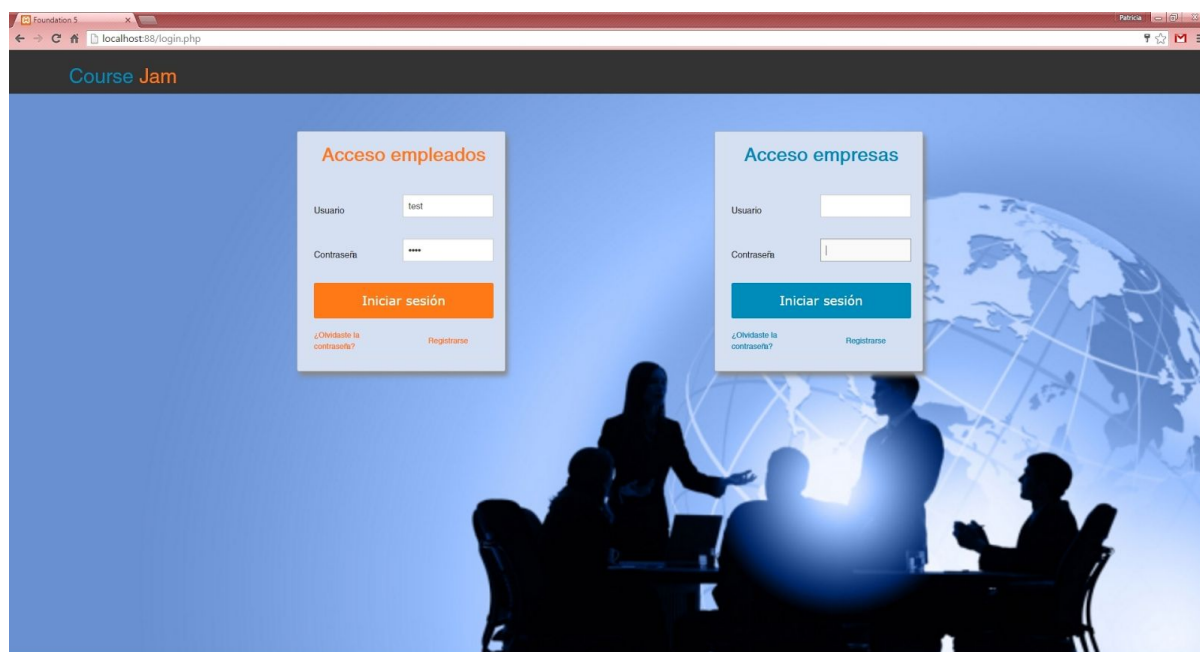


Figura 50: Página web final - Pantalla de inicio de sesión



Figura 51: Mockup para comparación con el resultado final

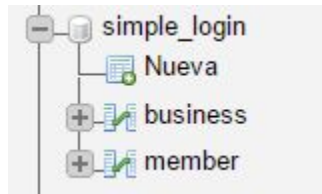


Figura 52: Página web final - Tablas de la base de datos

+ Opciones		bus_id	busname	buspw	bus_size	bus_sector	bus_pref	bus_picture
← T →		1	empresa	empresa	VSE	Servicios y tecnologías de la información	Nuevas tecnologías	[BLOB - 64 KB]
Editar Copiar Borrar								

Figura 53: Página web final - Tabla business con los datos de empresa

+ Opciones		mem_id	username	password	fname	lname	occupation	degree	picture	MOOCpref	personal	environment
← T →		4	test	test	Patricia	Rodriguez	programadora	ingeniera	[BLOB - 64 KB]	Bases de datos	Programacion web y diseño grafico	Online
Editar Copiar Borrar												

Figura 54: Página web final - Tabla 'member' con los datos de usuario

En la parte inferior de cada formulario se encuentran dos enlaces:

El enlace de la izquierda se trata de un servicio para restablecer la contraseña en caso de pérdida, que no se encuentra implementado. Para implementarlo, se debería añadir un nuevo campo a las tablas de la base de datos, para que el usuario pudiera introducir su dirección de correo electrónico. También se podría considerar como mejora de la web la opción de permitir el inicio de sesión tanto con un nombre de usuario como con la dirección de correo electrónico.

El enlace de la derecha redirige al usuario a la pantalla de registro, como usuario (figura 55) o como empresa (figura 56).

La página de registro como usuario contiene los campos definidos para el perfil de usuario en el punto 3.3.2, siendo obligatorio rellenar los campos 'Nombre', 'Apellidos', 'Nombre de usuario' y 'Contraseña'. En la figura 57 se puede ver el formulario completo de registro. En caso de que alguno de estos campos no contenga información aparecerá un aviso de que esa información es necesaria para continuar (figura 58).

Asimismo, la figura 59 contiene el formulario de registro como empresa, con los campos provistos en el punto 3.2.2, siendo obligatorio rellenar los campos 'Nombre de la empresa', 'Tamaño de la empresa', 'Sector o Industria' y 'Contraseña', apareciendo un aviso en caso de que alguno de estos campos no contenga información.

Foundation 5

localhost:88/register.php

Course Jam

Formulario de registro como nuevo usuario

Nombre:

Apellidos:

Profesión:

Grado académico:

Imagen:  Ningún archivo seleccionado

Preferencia del entorno de aprendizaje:

Preferencias de MOOCs e intereses:

Nombre de usuario:

Contraseña:

Figura 55: Página web final - Pantalla de registro de usuario

Foundation 5

localhost:88/registerEmpresa.php

Course Jam

Formulario de registro como empresa

Nombre de la empresa:

Tamaño de la empresa:

Sector o industria:

Preferencias e intereses:

Logo de la empresa:  Ningún archivo seleccionado

Contraseña:

Figura 56: Página web final - Pantalla de registro de empresa

### Formulario de registro como nuevo usuario

Nombre:

Apellidos:

Profesión:

Grado académico:

Imagen:

Seleccionar archivo

Ningún archivo seleccionado

Preferencia del  
entorno de  
aprendizaje:

Online



Preferencias de  
MOOCs e  
intereses:

Nombre de  
usuario:

Contraseña:

Enviar

Figura 57: Página web final - Formulario de registro de usuario

The image shows a web browser window with a registration form. A modal alert box is open, titled "localhost:88 dice:", with the text "Introduce tu nombre" and an "Aceptar" button. The form in the background includes fields for "Ape" (likely Apellido), "Prof" (likely Profesión), "Grado académico:", "Imagen:" with a "Seleccionar archivo" button and the text "Ningún archivo seleccionado", and "Preferencia del".

localhost:88 dice: x

Ape Introduce tu nombre

Prof

Grado académico:

Imagen: Seleccionar archivo Ningún archivo seleccionado

Preferencia del

*Figura 58:* Página web final - Alerta de formulario de registro de usuario incompleto

**Formulario de registro como empresa**

Nombre de la empresa:

Tamaño de la empresa:

Sector o industria:

Preferencias e intereses:

Logo de la empresa:  Ningún archivo seleccionado

Contraseña:

Figura 59: Página web final - Formulario de registro de empresa

Al rellenar correctamente los campos de uno de los registros, esta información pasa a formar parte de la base de datos, creando un nuevo usuario o empresa con la información registrada. Al introducir en la pantalla de inicio de sesión información que se corresponda con la que se encuentra en la tabla de miembros de la base de datos se accede a la página principal para empleados (figura 60 y figura 61). En esta página se encuentra el buscador de cursos creado en el punto 5.2, seguido por un slider donde se podría ver los cursos que está realizando el usuario, característica a implementar en el futuro. Para ello, habría que crear un nuevo campo en la base de datos, en el cual se guarde esta información.

La parte del buscador se diferencia ligeramente de lo esperado en la creación del diseño, ya que se trata de un iframe, esto es, un elemento html que permite insertar un documento HTML, en este caso el renderizador desde la página de OpenSearchServer dentro del documento HTML principal, que sería en este caso “search.php” [51]. No se ha podido añadir



un botón para iniciar la búsqueda, si no que solamente se dispone de la barra de búsqueda, ya que las opciones para modificar este iframe son muy limitadas, pudiendo solamente editar el css y algunas opciones html.

Para crear el slider se ha utilizado *slick carousel* [52], un proyecto compatible con foundation muy fácil de utilizar y con muchas opciones distintas disponibles. Las características de este slider son las siguientes:

- Totalmente responsivo. Escala con su contenedor.
- Ajustes separados para cada breakpoint
- Utiliza CSS3 cuando está disponible. Cuando no lo está sigue siendo completamente funcional.
- Se puede utilizar desplazando el dedo sobre la pantalla (swipe). Esta característica también se puede deshabilitar si se prefiere.
- También permite arrastrar el ratón sobre el escritorio para realizar el movimiento.
- Bucle infinito.
- Totalmente accesible utilizando la navegación de las teclas de flecha.
- Se puede añadir, eliminar, filtrar y eliminar filtrado de diapositivas.
- Reproducción automática, puntos, flechas, retrollamadas, etc ...

Slick no sólo funciona para imágenes, si no para bloques completos, pudiendo contener la información que se desee. Este slider responde a las necesidades de la página web para mostrar tanto los cursos que está realizando el usuario como los cursos finalizados.

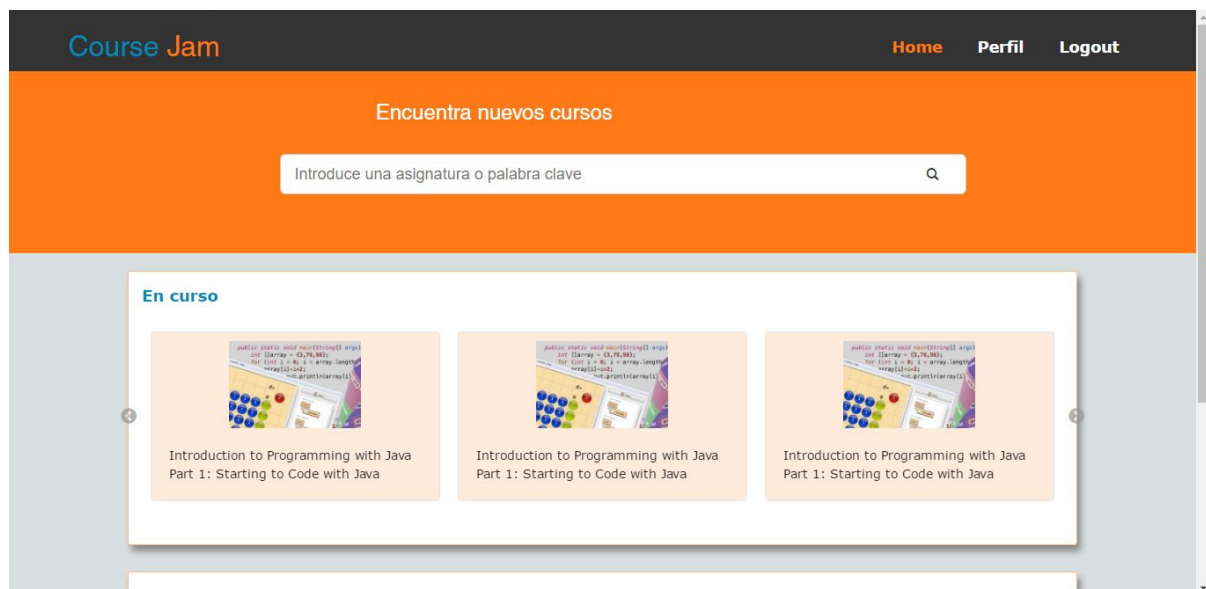


Figura 60: Página web final - Página inicial: Buscador de cursos y cursos en proceso

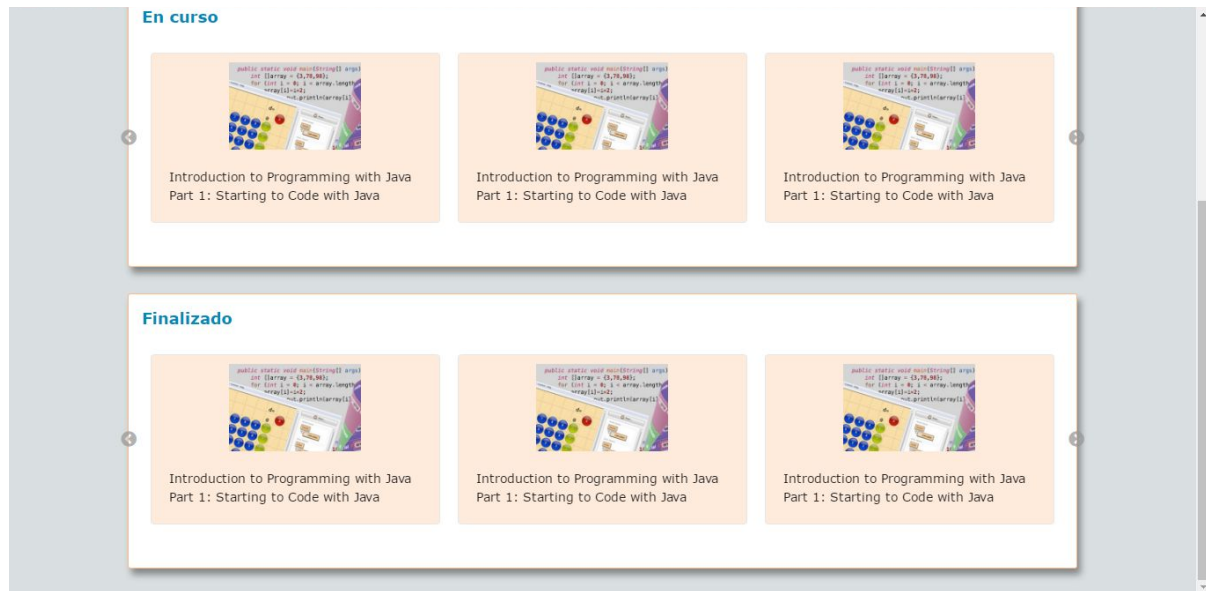


Figura 61: Página web final - Página inicial: Cursos en proceso y finalizados

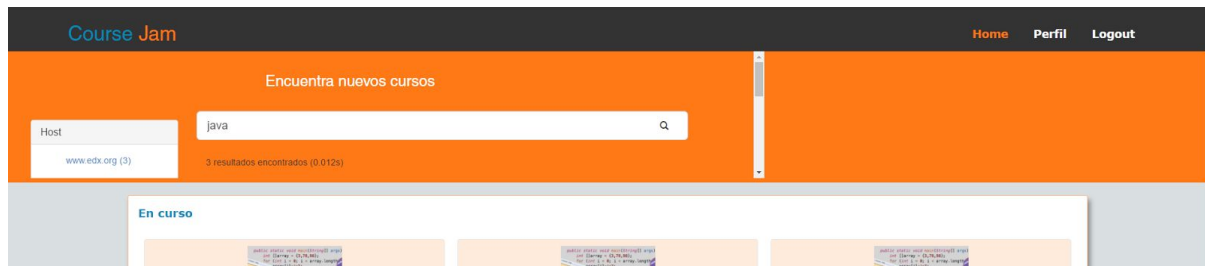


Figura 62: Mockup de la página principal para comparación con el resultado final

En la esquina superior derecha se encuentran tres pestañas diferentes, marcado en naranja el lugar donde se encuentra el usuario actualmente, en este caso en la página principal. La

segunda opción llevará al usuario a la página con el perfil personal, tal y como se desarrolla más adelante, y la tercera opción “Logout” terminará la sesión actual y devolverá al usuario a la página de inicio de sesión.

Al probar la eficacia del buscador utilizando esta configuración surge un problema: La búsqueda de cursos se realiza en la misma página en lugar de proveer una nueva dirección URL para visualizar los resultados. Por tanto, la visualización de la página con los resultados de búsqueda será errónea, como se muestra en la figura 63.



*Figura 63: Problemas en la pantalla de búsqueda de cursos*

Para arreglar este problema existen varias soluciones disponibles.

La primera solución consiste en utilizar Javascript para cambiar el tamaño del div donde se encuentra el iframe o enviar al usuario a una nueva ventana cuando se pulsa la tecla “ENTER” para realizar la búsqueda. Esta solución es poco robusta, ya que si el usuario fuera a pulsar esa tecla por cualquier otra razón también se cambiaría la vista, pudiendo resultar confuso para el usuario en algunas ocasiones. Si en lugar de un iframe se pudiera acceder al código del buscador esta opción podría ser implementada con mayor facilidad, ya que se podría disponer de un botón para comenzar la búsqueda, tal y como aparecía en el diseño inicial, y utilizar Javascript como se sugería anteriormente para modificar la página web.

La segunda opción se trata de simplificar la página: Como se va a mostrar posteriormente, los cursos en proceso y cursos finalizados que aparecen a continuación del buscador se encuentran también en el perfil del usuario, por tanto se podría eliminar esta información de la página de búsqueda para que no sea redundante, y de esta manera se solucionaría también el problema actual.

Al implementar la segunda solución, la página de búsqueda resultaría de la manera que se observa en la figura 64. Al introducir una palabra clave aparecerán opciones de autocompletado (figura 65) y después de introducir la frase que se desea buscar y presionar la tecla ENTER se mostrarán los resultados de la búsqueda, tal y como aparece en la figura 66.



Figura 64: Página web final - Página inicial: Buscador de cursos arreglado

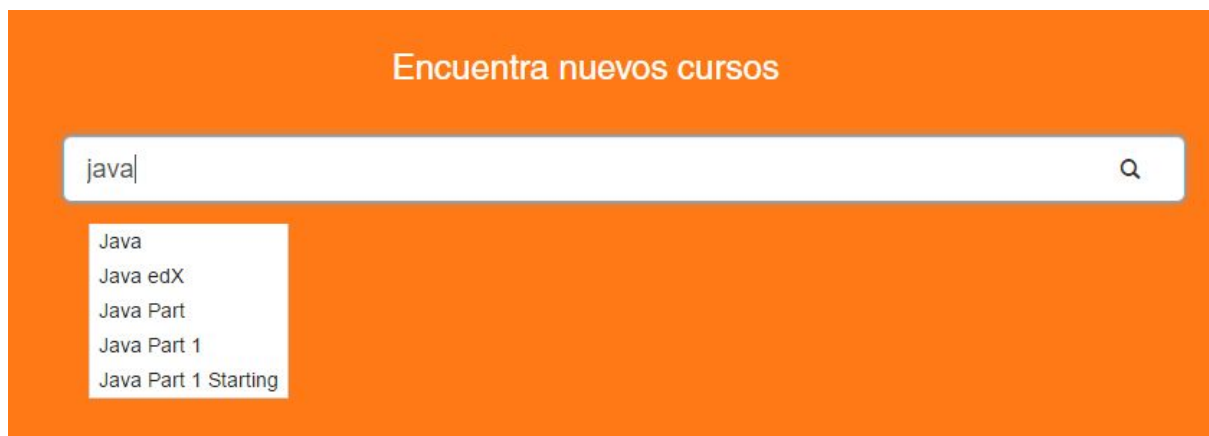


Figura 65: Página web final - Página inicial: Autocompletado en las búsquedas

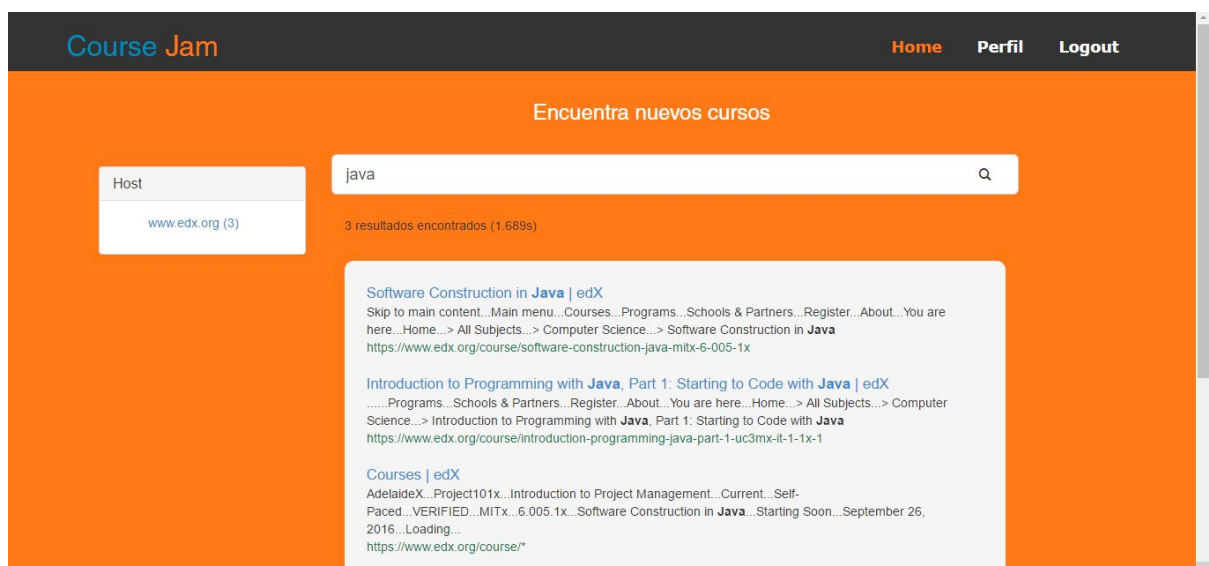


Figura 66: Página web final - Página inicial: Resultados de búsqueda

Si el usuario hace click en la pestaña “Perfil” podrá acceder a la ventana que contiene el perfil del usuario (figura 67), además de mostrar de nuevo los cursos en proceso y los finalizados (figura 68).

Aunque se contiene un campo en la tabla de la base de datos para la imagen, no se ha podido insertar una imagen desde archivo, lo cual queda como trabajo futuro. Para realizarlo, habría que crear un script php separado para recoger la imagen de la base de datos y de esta manera poder mostrarla [53]. También queda por desarrollar la posibilidad de editar los campos del perfil, para que el usuario pueda editar los valores y rellenar el perfil.

Junto al perfil del empleado se encuentra un hueco disponible para añadir las características de alertas y notificaciones, opciones que no se encuentran dentro del ámbito de este trabajo.

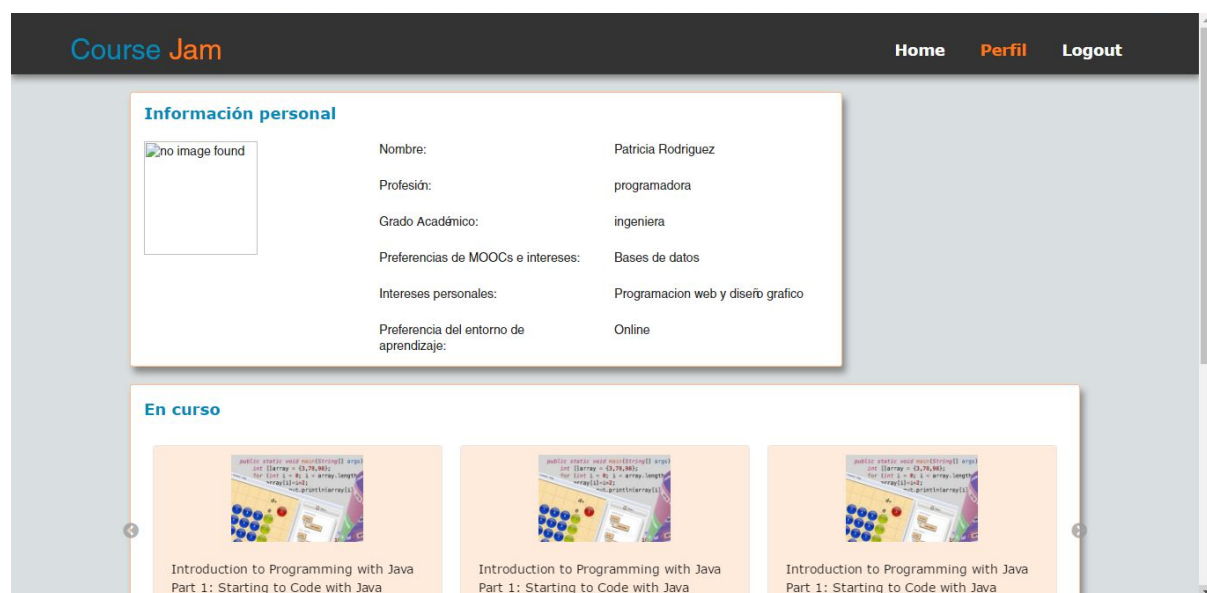


Figura 67: Página web final - Perfil del usuario y cursos en proceso



Figura 68: Página web final - Cursos en proceso y finalizados



Figura 69: Mockup de la página de perfil de usuario para comparación con el resultado final

Retrocediendo a la pantalla de inicio de sesión, si el usuario introduce información referente a una empresa en los campos correspondientes puede acceder al perfil de empresa que se observa en la figura 70 y figura 71.

De manera análoga a como ocurre en el perfil de usuario, se ha dejado una zona en blanco a la derecha del perfil de la empresa, para añadir próximamente las opciones de añadir empleados, que se visualizarán en el primer slider debajo del perfil, o añadir cursos recomendados, que se encuentran en el slider posterior. También queda como trabajo futuro la opción de editar el perfil de la empresa y poder añadir una imagen.

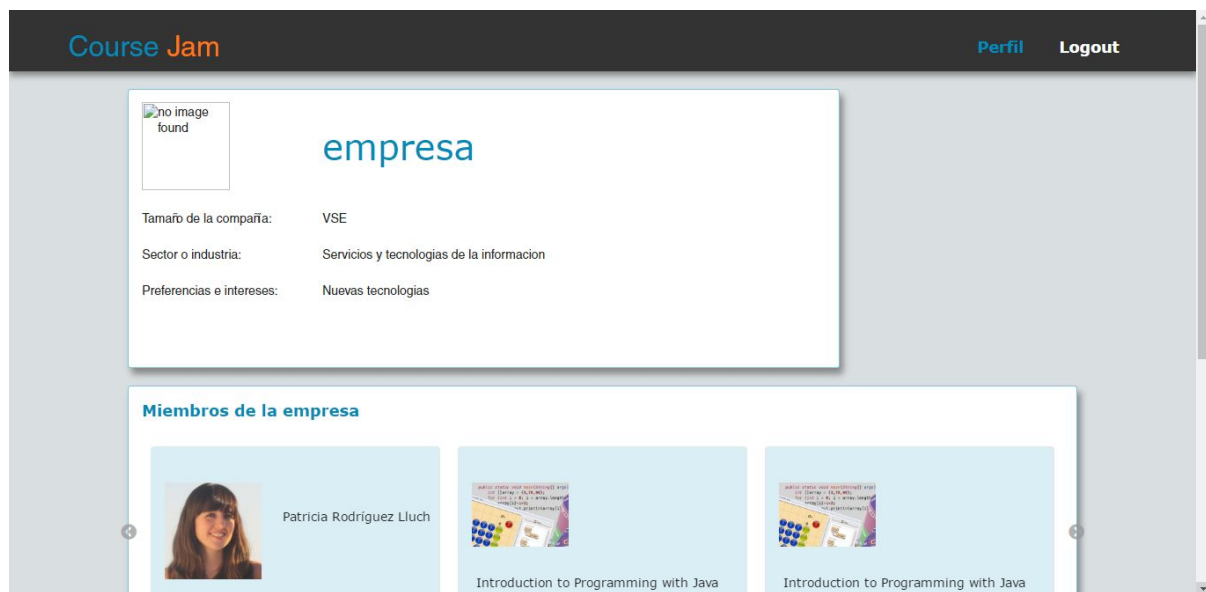


Figura 70: Página web final - Perfil de empresa y miembros



Figura 71: Página web final - Miembros de la empresa y cursos recomendados





Figura 72: Mockup de la página de perfil de empresa para comparación con el resultado final



## 6.2 Conclusiones

In this work, a new MOOC aggregator has started taking shape, oriented to VSEs as contrary to the already existing MOOC aggregators.

- First of all the currently most popular MOOC aggregators web pages (Class Central, CourseBuffet and CourseTalk) were investigated, compared and analyzed. The analysis concluded that all three of them are more focused on students and the academic world in general. Another common factor among all three Web applications are the MOOC searching capacities. The particular key points of the analyzed MOOC-Aggregators are:
  - Class Central: Notification & Alerts
  - CourseBuffet: Presentation & Resources, Benchmarking
  - CourseTalk: Rating, Social
- Based on these analysis, a set of characteristics that a MOOC aggregator should have are shown, in order to decide the focus of the design of a VSEs-oriented MOOC-Aggregator Web application for supporting ETSD in small settings. The proposed characteristics are shown here below:
  - VSE-oriented
  - Searching
  - Benchmarking & Rating
  - Customization & personalization
  - Notification & Alerts
  - Social Web
- The author chose VSE-oriented, Searching and Customization & personalization as the main keys for the scope of this project.
- The front-end of the webpage was designed using mockups and then developed, choosing Foundation as the framework and using HTML, CSS, Javascript and PHP as programming languages.
- A database has been created, containing the information of the users and business that create an account on the web.
- A search tool has been developed, so users can search for courses from different MOOC aggregators.
- Overall, the purposes of the author in order to start the creation of the webpage were fulfilled, although some elements need to be polished in order to create a fully functional MOOC aggregator Web application.

## 6.3 Trabajo futuro

Además de las mejoras ya propuestas durante el transcurso del trabajo, hay otras ideas que sería de utilidad implementar en la plataforma web.

En vista de los resultados ofrecidos por el buscador de MOOC creado con OpenSearchServer, se podría realizar un seguimiento de las búsquedas de los usuarios, para poder refinar los resultados. También se debe estudiar la opción de crear desde cero un buscador, lo cual tendría una parte compleja de programación y resultaría en un nivel más básico de opciones de búsqueda, pero sin embargo aportaría una mayor flexibilidad a la página web, sin dependencias externas.

Por otra parte, es necesario conectar los cursos encontrados a los perfiles de usuario, para que los empleados puedan acceder a los cursos que quieren realizar y puedan ver el progreso del mismo. En este mismo marco también se engloba la implementación de un sistema de notificaciones y alertas para avisar a los usuarios de los nuevos cursos que comienzan. Asimismo, la parte web orientada a empresa debe ser capaz de recoger la información de sus empleados y de los cursos que quiere ofrecer a los mismos.

# Lista de referencias

1. Sevillano, María Luisa (Coord.) (2002). Nuevas tecnologías, medios de comunicación y educación. Madrid: CCS. p. 15.
2. [https://es.wikipedia.org/wiki/Educación\\_en\\_línea](https://es.wikipedia.org/wiki/Educación_en_línea)
3. Westhead, P., Storey, D.: Management Training and Small Firm Performance: Why is the Link So Weak? *Int. Small Bus. J.* 14, 13–24 (1996).
4. Laporte, C., O'Connor, R., Fanmuy, G.: International systems and software engineering standards for very small entities. *CrossTalk J. Def. Softw. Eng.* 26, 28–33 (2013).
5. Moreno Campos, E., Sanchez-Gordón, M.-L., Colomo-Palacios, R., de Amescua Seco, A.: Towards Measuring the Impact of the ISO/IEC 29110 Standard: A Systematic Review. In: *Systems, Software and Services Process Improvement*. pp. 1–12. Springer (2014).
6. Gibb, A.A.: Small Firms' Training and Competitiveness. Building Upon the Small business as a Learning Organisation. *Int. Small Bus. J.* 15, 13–29 (1997).
7. Roy, A., Raymond, L.: Meeting the training needs of SMEs: is e-learning a solution? *Electron. J. E-Learn.* 6, 89–98 (2008).
8. Luján Mora, S., others: Los MOOCs, ¿pueden interesar a Recur. Hum. Relac. Laborales. 80, 48–52 (2013).
9. Hamburg, I., O. Brien, E., Engert, S.: Engaging SMEs in Cooperation and New Forms of Learning. *Comput. Inf. Sci.* 7, 1–9 (2014).
10. Moreno Campos, E. J., Evaluation of MOOC-Aggregators for supporting Education, Training and Staff Development in VSEs (2015).
11. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 107–117 (1998).
12. An, B., Qu, T., Qi, H., Qu, T.: Chinese MOOC Search Engine. In: Wang, H., Qi, H., Che, W., Qiu, Z., Kong, L., Han, Z., Lin, J., and Lu, Z. (eds.) *Intelligent Computation in Big Data Era*. pp. 453–458. Springer Berlin Heidelberg (2015).
13. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., and Nejdl, W. (eds.) *The 79 Adaptive Web: Methods and Strategies of Web Personalization*. pp. 3–53. Springer-Verlag, Berlin, Heidelberg (2007).
14. Brusilovsky, P., Kobsa, A., Nejdl, W.: *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer Science & Business Media (2007).
15. Mehtaa, P., Parekh, B., Modi, K., Solanki, P.: Web Personalization Using Web Mining: Concept and Research Issue. *Int. J. Inf. Educ. Technol.* 2, (2012).
16. Eirinaki, M., Vazirgiannis, M.: Web Mining for Web Personalization. *ACM Trans Internet Technol.* 3, 1–27 (2003).

17. Appelquist, D., Brickley, D., Carvahlo, M., Iannella, R., Passant, A., Perey, C., Story, H.: A Standards-based, Open and Privacy-aware Social Web. W3C (2010).
18. <http://www.awwwards.com/what-are-frameworks-22-best-responsive-css-frameworks-for-web-design.html>
19. <http://getskeleton.com/>
20. <http://designinstruct.com/roundups/html5-frameworks/>
21. <http://www.vermilion.com/responsive-comparison/>
22. <http://webdesign.tutsplus.com/es/articles/a-quick-guide-to-choosing-the-best-framework-for-you--cms-23575>
23. <http://getbootstrap.com/>
24. <http://foundation.zurb.com/>
25. Matt Ward: One Unexplored Reason Why Photoshop Won't Be Dethroned Anytime Soon, MyInkBlog
26. <http://www.hosting.vt.edu/tutorials/phpmysql/#about>
27. Susan de Sousa's - My project Manager expert. Different Types of Databases
28. [https://es.wikipedia.org/wiki/Base\\_de\\_datos\\_relacional](https://es.wikipedia.org/wiki/Base_de_datos_relacional)
29. <http://tecnocacharrero.blogspot.com.es/2009/10/mtop-monitoreo-de-carga-en-mysql.html>
30. <https://es.wikipedia.org/wiki/MySQL>
31. «License». PostgreSQL Global Development Group. Consultado el 24 de agosto de 2015.
32. <https://es.wikipedia.org/wiki/PostgreSQL#Caracter.C3.ADsticas>
33. <https://www.phpmyadmin.net/>
34. Windows Server Administration Fundamentals. Microsoft Official Academic Course. 111 River Street, Hoboken, NJ 07030: John Wiley & Sons. 2011. pp. 2–3. ISBN 978-0-470-90182-3.
35. Comer, Douglas E.; Stevens, David L. (1993). Vol III: Client-Server Programming and Applications. Internetworking with TCP/IP. Department of Computer Sciences, Purdue University, West Lafayette, IN 479: Prentice Hall. pp. 11d. ISBN 0-13-474222-2.
36. [https://en.wikipedia.org/wiki/Server\\_\(computing\)](https://en.wikipedia.org/wiki/Server_(computing))
37. <https://es.wikipedia.org/wiki/Escalabilidad>
38. [http://blog.gainlo.co/index.php/2016/06/29/build-web-crawler/?utm\\_source=quora&utm\\_medium=How+can+I+build+a+web+crawler+from+scratch%3F&utm\\_campaign=quora](http://blog.gainlo.co/index.php/2016/06/29/build-web-crawler/?utm_source=quora&utm_medium=How+can+I+build+a+web+crawler+from+scratch%3F&utm_campaign=quora)
39. <http://www.robotstxt.org/>
40. [https://es.wikipedia.org/wiki/Deduplicaci%C3%B3n\\_de\\_datos](https://es.wikipedia.org/wiki/Deduplicaci%C3%B3n_de_datos)
41. [https://en.wikipedia.org/wiki/Bloom\\_filter](https://en.wikipedia.org/wiki/Bloom_filter)
42. <http://www.openwebspider.org/>
43. <https://www.webcrawling.net/>

44. <https://zscraper.wordpress.com/>
45. <http://www.opensearchserver.com/>
46. <https://en.wikipedia.org/wiki/OpenSearchServer>
47. <http://www.opensearchserver.com/documentation/tutorials/functionalities.md>
48. [https://es.wikipedia.org/wiki/Servidor\\_proxy](https://es.wikipedia.org/wiki/Servidor_proxy)
49. <http://es.ccm.net/contents/649-formatos-y-extensiones-de-archivos>
50. [https://es.wikipedia.org/wiki/Token\\_\(inform%C3%A1tica\)](https://es.wikipedia.org/wiki/Token_(inform%C3%A1tica))
51. <https://es.wikipedia.org/wiki/Iframe>
52. <http://kenwheeler.github.io/slick/>
53. <http://stackoverflow.com/questions/7793009/how-to-retrieve-images-from-mysql-data-base-and-display-in-an-html-tag>.

# Lista de referencias de figuras

- 2.1 Moreno Campos, E. J. (2015). Evaluation of MOOC-Aggregators for supporting Education, Training and Staff Development in VSEs [Figura].
- 2.2 <http://www.opensearchserver.com/documentation/tutorials/functionalities.md> [Figura]